How Much Do I Argue Like You? Comparing Attitudes in Argumentation and Derived Applications

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Markus Brenneis

geboren in

Neuss

Düsseldorf, April 2021

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Martin Mauve
2. Prof. Stefan Harmeling

Tag der mündlichen Prüfung: 23.08.2021

Nur weil ihr die gleiche Meinung habt, heißt das nicht, dass ihr auf derselben Seite seid.

Lian Remme, 7. August 2020

Abstract

Many people take part in online discussions, state their opinions on controversial topics, and bring forward their arguments. Sometimes, you ask yourself how similar your opinion is to the opinion of another participant of the discussion. For example, you might be reading the attitudes of political parties towards current political issues, and you ask yourself which party to vote for. But to undertake this kind of comparison, you need some means to calculate the similarities of attitudes in an argumentation.

In this work, we delve into the issue of determining the similarities of individual views in an argumentation. We present a theoretical model to capture different opinions and arguments in argumentation contexts and develop a pseudometric for calculating the (dis)similarity between the views of two participants. Furthermore, we investigate how to make sure that such a distance measure yields intuitive results by looking at an empirical study where we collected human baseline results for argumentation similarity assessments. We propose different distance functions and study which best match human intuition and where the functions have limitations.

Once we have the theoretical means to compare attitudes in argumentations, we examine two possible use cases. First, we explore how to achieve a clearer view in online discussion platforms with numerous arguments by pre-filtering arguments using neighborhood-based collaborative filtering. Our new argumentation platform *deliberate* includes such a filtering algorithm which uses our pseudometric for calculating the similarity of users based on their attitudes in the argumentation. We expound on results from an experiment with *deliberate*, where the influence of different filtering algorithms on the formation of opinion was studied. Moreover, we present our argumentation dataset for evaluating argument recommender systems which comprises several hundred user profiles.

As a second use case, we introduce our argument-based Voting Advice Application (VAA) ArgVote, which computes the similarity of political views of parties and voters not only based on their opinion concerning central theses, but also considering their arguments. Although we could not demonstrate that our new matching algorithm based on our pseudometric was more accurate than the algorithm in classical VAAs, we were, nevertheless, able to show positive effects of our argument-based system on the understanding of political issues. The dataset containing the user profiles of our study participants is provided to improve the matching algorithms in future work. We subsequently present our idea for a VAA chat bot to address some issues with ArgVote which our experiment revealed.

Our work lays the foundation for further exciting applications in the context of argumentations, for instance the clustering of voters, an automatic finding of compromises, or escaping filter bubbles. The impact of the developed methods, systems, and different user interfaces on opinion formation or political interest can be further researched in larger empirical studies.

Zusammenfassung

Viele Menschen nehmen an Onlinediskussionen teil, sagen ihre Meinungen zu kontroversen Themen und bringen ihre Argumente ein. Manchmal fragt man sich, wie ähnlich die eigene Meinung zu der Meinung anderer Diskussionsteilnehmer:innen ist. Man könnte z. B. gerade die Einstellungen von Parteien zu aktuellen politischen Themen lesen und sich fragen, welche Partei man wählen soll. Aber um diese Art von Vergleich vornehmen zu können, braucht es eine Methode, die Ähnlichkeit von Einstellungen in einer Argumentation zu berechnen.

In dieser Arbeit beschäftigen wir uns damit, die Ähnlichkeit individueller Ansichten in einer Argumentation zu bestimmen. Wir präsentieren ein theoretisches Modell zum Festhalten verschiedener Meinungen und Argumente und entwickeln eine Pseudometrik zur Berechnung des Abstands zwischen den Ansichten zweier Personen. Ferner untersuchen wir, wie man sicherstellen kann, dass ein solches Distanzmaß intuitive Ergebnisse liefert, indem wir unsere empirische Studie ansehen, in der menschliche Referenzwerte für die Ähnlichkeitsbewertung von Argumentationen gesammelt wurden. Wir schlagen verschiedene Distanzfunktionen vor und betrachten, welche am besten mit menschlicher Intuition übereinstimmen und wo ihre Grenzen liegen.

Ausgestattet mit diesen theoretischen Mitteln untersuchen wir zwei mögliche Anwendungsfälle. Zuerst erkunden wir, wie eine übersichtlichere Ansicht vieler Argumente auf Diskussionsplattformen durch eine Vorfilterung von Argumenten mit Nachbarschafts-basiertem Collaborative Filtering erreicht werden kann. Unsere neue Argumentationsplattform *deliberate* beinhaltet einen solchen Algorithmus, der mit unserer Pseudometrik die Ähnlichkeit von Teilnehmer:innen basierend auf deren Einstellungen in der Argumentation berechnet. Wir erläutern Ergebnisse von einem Experiment mit *deliberate* zum Einfluss verschiedener Filteralgorithmen auf die Meinungsbildung. Außerdem präsentieren wir unseren Argumentationsdatensatz zur Evaluation von Recommender-Systemen für Argumente, der hunderte Benutzerprofile beinhaltet.

Als zweite Anwendung stellen wir unsere Argument-basierte Voting Advice Application (VAA) ArgVote vor, die die Ähnlichkeit von politischen Ansichten von Parteien und Wähler:innen nicht nur basierend auf deren Meinung zu zentralen Thesen, sondern unter Berücksichtigung von Argumenten berechnet. Wir konnten zwar nicht zeigen, dass unser Vergleichsalgorithmus, der auf unserer Pseudometrik basiert, akkuratere Ergebnisse liefert als der klassische Algorithmus, jedoch stellten wir einen positiven Effekt unseres Systems auf das Verständnis politischer Themen fest. Der Datensatz mit den Profilen der Studienteilnehmer:innen steht zur Verfügung, um den Vergleichsalgorithmus in Zukunft verbessern zu können. Zum Adressieren ein paar der festgestellten Probleme mit ArgVote präsentieren wir unsere Idee für einen VAA-Chatbot.

Wir legen die Grundlage für weitere Anwendungen im Kontext von Argumentationen, beispielsweise Clustering von Wähler:innen, automatische Konsensfindung und Entkommen aus Filterblasen. Die Auswirkungen der entwickelten Methoden, Systeme und neuer Benutzeroberflächen auf Meinungsbildung und politisches Interesse können noch weiter untersucht werden.

Acknowledgements

There are many people I have to thank for making this thesis possible. I want to thank you all in more or less chronological order, hopefully not forgetting anyone.

First, I would like to thank Andre and my colleague Alexander, who causally contributed to me getting a PhD position at the computer networks group—I will not forget the moment Andre showed up at the Fachschaft to make me aware of the open position.

I am very grateful to Martin for giving me the opportunity to work in his group. He helped me to establish valuable contacts with interdisciplinary research groups and was always—even on vacation—interested in my work and quick to give feedback.

Thanks to my office mate Björn, with whom I could talk about crazy ideas on the whiteboard, who had to bear my simple programming questions about Fulcro, and who ensured a regular tea supply. And thanks for the copy of the robot.

Thank you, Thomas, Sabine, and Angela, for helping out with all the technical and organizational tasks. You were always fast to help when software problems or administration questions arose.

I also have to thank Stefan for being my co-supervisor, and providing feedback for my work. Thanks to you, Maike, my colleagues, and our Hiwis for many fruitful discussions.

Moreover, it was a great experience to be part of the NRW Forschungskolleg Onlinepartizipation and the UPEKI research group. Many interdisciplinary discussions have broadened my horizon, e.g. regarding topics like confirmation bias and cognitive dissonance. Special thanks go to Ole for the close planning and evaluation of surveys, as well as the colleges from the communication and political science for supporting my first steps with empirical studies.

I am thankful for many great years with organizing lectures and teaching, especially for Computer Architecture with Janine. It taught me good organization of lectures and inspired new concepts for online and classroom teaching.

Thanks to a little guy called SARS-CoV-2, I could not physically take part in any international conference so far. At least the environment is thankful.

Finally, I would like to thank everyone who has proofread parts of my dissertation, and my family and friends who have supported me throughout my studies. Hopefully, we will have many more Nichtenttäuschungen and other celebrations together (without physical distancing).

Contents

Chapter 1

Introduction

With the rise of the Internet, exchanging arguments and sharing one's own opinion with a wide audience has become very simple. Individual persons as well as organizations can bring forward their arguments and react to the statements of others. Stating one's arguments to a position is used, for instance, by political parties to back their opinions, or on smaller scales when discussing how to spend money at a university (Ebbinghaus, 2019). But reading through the flood of available information is not easy. How can we focus on arguments which are most relevant to ourselves? Is it possible to find people or organizations with similar or different arguments and stances than oneself? But how can you measure how similar two people argue?

In this work, we delve into the subject of calculating how close two participants in an argumentation argue. We have a look at how a distance measure for this task can help pre-filtering arguments presented to users and applications for calculating the similarity between political parties and voters.

Before we can measure any distance, we have to get the arguments in some structured way which is not just plain text. Argument mining is one approach to extract structured argument information from natural language texts (Lawrence and Reed, 2020). With those methods, for example, automatic analysis of arguments in online participation projects (Liebeck et al., 2016) or real-time search of arguments during a pandemic (Daxenberger and Gurevych, 2020) can be implemented. But it is still not feasible for most people to read and understand more than 100 arguments about the danger of a novel virus or the implications of wearing face masks, even if we assume a perfectly structured and non-redundant presentation.

There are some solutions which address the problem of having to read through all arguments. The *Dialog-Based Argumentation System* (D-BAS) by Krauthoff et al. (2017) presents arguments one-by-one in a chat-like fashion; but the system focuses on countering my current stance and fails to give a complete overview of a discussion. Chalaguine and Hunter (2020), on the other hand, developed a chat bot which tries to understand my concerns and react to them.

Ideally, we need a system which knows a user's attitudes, and then can select arguments which are interesting and relevant to a user. This system should not present arguments the user already knows or does not accept, but introduce new information to assist their formation of opinion. When the system is familiar with a user's stances, it could also tell which political parties or organizations are most in line with their attitudes. That way, an argumentation tool could assist in making voting decisions for the next general election.

But in order to build such a system, we must first create some sort of sensible pre-filtering algorithm. One approach could be collaborative filtering using nearest-neighbor algorithms, which somehow have to measure the similarity of users. Such a similarity measure is also needed for calculating the user–party distance for voting recommendations, so our application scenarios call for an answer to the following question: *How can the similarity of attitudes of participants in an argumentation be calculated?*

1.1 Research Questions

We break this question of how to compare attitudes in argumentation down into several sub-questions. First, we concentrate on the theoretical foundations, and then move on to questions arising in applications of an argumentation distance function.

How can the pieces of information in real-world argumentations be represented mathematically? When we look at argumentations in the real world, people express their attitudes in different dimensions. They (partially) agree or disagree with arguments, attack or defend other people's arguments, give some issues higher priorities than others etc. If we want to compare such attitudes, we need a mathematical model to capture all those pieces of information. After that, a metric could be defined on that structure, which measures the distance between two instances.

How can we calculate that distance? Once we have argumentations and individual attitudes captured in a mathematical model, we must somehow calculate the distance between two instances of the model. In the best case, such a distance function should be a metric, fulfilling properties like the triangle inequality. This property allows re-using existing clustering or optimized nearest-neighbor algorithms, which expect a metric as input.

How can we be sure that our way of calculation is intuitive and can be understood? After having developed a distance function, one has to ask why it should be trusted. Does it yield results which are in line with human intuition? Can the calculated values be trusted, or do they appear random? A good evaluation method would be comparing the function's results with assessments of human beings. For this, one has to present different people various argumentation situations and ask for their similarity assessment. This human baseline can then be used to benchmark different possible metrics.

How can we collect the necessary information from users in real-world applications? A mathematical model and a validated metric are nice tools, but they are of no use if the model needs pieces of information not available in practice. We have to ensure that the information needs of our model can be met in reality. This means we must be able to build applications which can find out, for instance, how sure a user is about a statement in an argumentation, while still having an intuitive and enjoyable *user interface* (UI).

How does the distance function perform in practice? Finally, a distance function and a well-planned UI are not worth anything if they do not produce usable results when being applied to real use-cases. How does the distance function perform when used in a kind of argument recommender system, which can be used to pre-filter the flood of arguments on the Web? Is it suitable for comparing the attitudes of voters and political parties and capable of recommending parties to vote for? If such a system is used, how does it influence the formation of opinions? Do people enjoy it? As we can see, practical applications have many more related questions than whether the mathematical idea turns out useful. One also has to pay attention to sensible UIs, strategies for content generation, implications on the formation of opinions, possible manipulation, etc. In this work, we will mainly focus on the general acceptability of applications and how well the metric-related features work. Thus, we concentrate on testing whether the UIs we consider intuitive are also accepted by real users, and whether our distance function improves the applications' results.

1.2 Contributions

In this thesis, we tackle the questions presented above. We present our model of weighted argumentation graphs which can capture a person's attitudes in an argumentation. It includes individuals' assessments of argument and position relevance and degrees of belief in statements.

Based on this new model, we develop a pseudometric which calculates how close the argumentations of two persons are. This pseudometric considers the special properties of argumentations around positions, e.g. by giving arguments supporting/opposing other arguments a smaller influence than arguments directly for/against a position. To assure that the calculations of the pseudometric yield sensible results, we present a list of intuitive desiderata (i.e. expected properties) and prove that these are fulfilled by it.

Afterwards, we consider the question of how to be sure that those desiderata, and hence, the pseudometric, actually match human intuition and not only the assessment of domain experts. We present the design and results of an empirical study we have conducted, which had the goal to find out how average people assess the similarity of various argumentation scenarios. Through this study, we found some surprising results which were not in line with our anticipated assessment as domain experts. For instance, our expectation that a supporting and an opposing view are more dissimilar to each other than a supporting and a neutral view could not be confirmed.

We subsequently adapt different existing distance functions for comparing attitudes in argumentations and check which of them best match the human intuition results we collected before.

As it turns out, other distance functions which are less complex than our pseudometric can be suitable for simpler, flat argumentation contexts, but they lack desired properties concerning deeper argumentations.

Next, we show different applications for comparing attitudes in argumentations. We introduce an argument recommender system used in our new web application *deliberate*, which has been used as part of a larger, interdisciplinary research project. Furthermore, we contribute a dataset with more than 600 user profiles and 900 arguments, which can be used to assess the performance of such recommender systems. We see that our nearest-neighbor approach performed better than a simple baseline classifier, but there is still room for improvements.

Finally, we examine how well-suited our pseudometric is for a *Voting Advice Application* (VAA). We introduce a new kind of argument-based VAA, ArgVote and evaluate it in an empirical study with 60 participants. This study compared two versions of ArgVote, where political positions were shown with or without arguments, respectively. We discover that the exposure to arguments improved the understanding of different opinions, and people liked interacting with the version which included arguments. The dataset from our study can be used to further improve argument-based matching algorithms. Lastly, we consider a VAA chat bot as an alternative UI to reduce the perceived time spent within the VAA.

1.3 Outline of This Thesis

In the next chapter, we explain the basics of argumentation theory necessary for following this thesis. We also look at other existing applications to see where our new applications are placed in the field of argumentation.

Chapter 3 deals with the theoretical background of our pseudometric. We introduce the model of weighted argumentation graphs and define our pseudometric based on that model. Afterwards, we present an empirical study about the assessment of argumentation similarity, and evaluate different distance functions against this baseline.

In Chapter 4, we look at applications of distance functions for comparing attitudes in argumentations. We introduce our argumentation application *deliberate*, which uses collaborative filtering to make online argumentations clearer, as well as the research project *AI support for policy decisions* (UPEKI), which employed the application. The dataset which we gathered and its usefulness for evaluating argument recommender systems is expounded. Finally, we present our new VAA ArgVote, look at the results of an empirical study conducted with it, and consider how it can be improved.

Last of all, Chapter 5 summarizes our main results and points out open questions for future work.

Chapter 2

Preliminaries and Related Work

Argumentation has been an important part of human history, and thus, has undergone research for a long time and from different perspectives. In this chapter, we have a look at the basics of argumentation theory and introduce the most relevant terms which will be relevant for this thesis. After that theoretic view, we explore which applications have been built and are being developed based on argumentation theory.

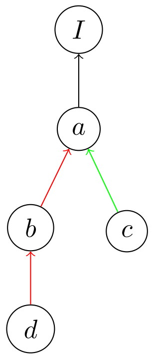
2.1 Argumentation Theory

The overall goal of argumentation theory is modeling of and applying algorithms to argumentations. An argumentation is an exchange of arguments like "We should *not* build more nuclear power plants because nuclear waste cannot be stored safely." (argument A). An argument comprises two separate statements (or propositions (Walton, 2009)): a premise ("Nuclear waste cannot be stored safely.") and a conclusion ("We should build more nuclear power plants.") (Freeman, 2011), which can be negated. Premise and conclusion are also referred to as evidence and claim, respectively (Booth et al., 2008). Note that some authors apply the word *argument* to a chain of arguments (Modgil and Prakken, 2014), whereas, in formal context, we only call a single premise–conclusion pair *argument*.

A premise can attack the conclusion, i.e. argue against it: "Nuclear waste *cannot* be stored safely." argues against "We should build more nuclear power plants." But a premise can also argue in favor of the conclusion, i.e. defend (or support) it, as in "Nuclear waste *can* be stored safely because it can be stored in deep geological repositories." (argument B).

There are three main kinds of attack relations in argumentation theory (Freeman, 2011; Modgil and Prakken, 2014; Pollock, 2001). When an argument targets the premise of another argument, like B and A in our example, B is said to undermine A. Another type of attack is a rebut, where the conclusion is attacked. For instance, A can be rebutted with "We should build more nuclear power plants because nuclear plants do not have any CO_2 emissions."

The last major attack relation is the undercut. An undercut does not attack a premise or a conclusion, but the relation of two statements, i.e. the validity of the argument. For example, "We should build more nuclear power plants because Party X suggested doing so." can be



I	Future of Energy Supply
a	We should build more nuclear power plants.
b	Nuclear waste cannot be stored safely.
c	Nuclear plants do not have any CO_2 emissions.
d	Nuclear waste can be stored in deep geological repositories.

Figure 2.1: An example of an argumentation graph. We see one position a, the attacking arguments (b, a) and (d, b), and the supporting argument (c, a).

undercut with "One's own opinion should not be based on who made a proposal." In particular, an undercut attack allows accepting both, the premise and the conclusion, but disagreeing with the argumentative relation.

Following the model of the *Issue-Based Information System* (IBIS) (Kunz and Rittel, 1970), there are special statements called positions, which are typically actionable items like "We should build more nuclear power plants." and are the anchor point of an argumentation. Therefore, they do not have a conclusion. Multiple positions can be about the same topic, which is referred to as issue.

For visualizations and further visual or theoretical analysis, the statements of an argumentation can be summarized in an argumentation graph, as depicted in Figure 2.1. The issue I is the root of the graph, the statements are the nodes, and the argument relations are the edges. In this kind of visualization, the positions look like "premises" of I, although they are semantically no premises.

A slightly different approach is taken by Dung (1995), who introduced the concept of abstract argumentation frameworks. An argumentation framework is a pair with a set of arguments and a binary attack relation, i.e. it is basically a graph with arguments as nodes and attacks as edges. Based on this popular definition of argumentation, different abstract tasks have been researched, i.a. the consistency or conflict-freeness of argumentations or determining sets of admissible arguments, which has applications e.g. in law (Bench-Capon and Modgil, 2009; Collenette et al., 2020).

Various other models for different application focuses have been developed based on Dung's model. Bipolar argumentation frameworks (Cayrol and Lagasquie-Schiex, 2005) add an additional support relation in addition to the attack relation. Other models add argument strengths or weights (Amgoud et al., 2017; Gordon and Walton, 2016; Hunter, 2013), which can then be used to calculate the most likely conclusions.

Hunter et al. (2020) presented epistemic graphs, which can model different degrees of argument acceptability. They also discussed the idea of having different views on a common graph by

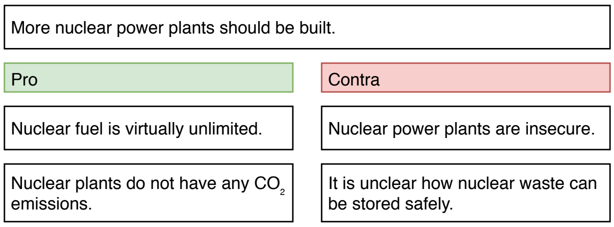


Figure 2.2: An example of a pro/con list, as they can be found in the UI of kialo, with four arguments for/against a statement.

different persons. For example, Alice might not believe that "nuclear waste can be stored in deep geological repositories," which can in turn influence her personal opinion on the position about building more nuclear power plants. We will present a similar model in Section 3.1.

2.2 Applications Built Around Argumentation

Based on the theoretic foundations of argumentations, different applications have been developed. In this thesis, we focus on applications aimed at end-users without domain knowledge, not on applications like Carneades (Gordon, 2013) and Neva (Yang et al., 2020) whose goal is assisting research about argumentation frameworks.

Extracting arguments and their structure from natural language texts is the goal of argument mining (Lawrence and Reed, 2020). When such an extraction recreates an argumentation graph, it can be used, for instance, for automatic analysis of online participation projects (Liebeck et al., 2016). As this automatic transformation is not perfect, programs which directly collect structured premise–conclusion relations have been developed.

On the web platform kialo[1], people can exchange arguments on different topics. With currently more than 2 million users and 13,000 topics, this is a very popular website for the structured exchange of arguments. The arguments added by the users are presented in pro/contra lists, similar to the view presented in Figure 2.2. Supporting and attacking arguments can be added to other arguments, resulting in an argumentation tree. Arguments can be commented on, and each user can rate an argument's impact.

A similar, but less popular application is ReasonScore[2], where arguments can be added in a tree-like structure. When adding an argument, it must either affect the confidence or relevance of the conclusion. The added argument then automatically influences the calculated degree of confidence of its conclusion. This way, it is possible to compute how much and why a certain position should be adopted.

A completely different approach is taken by the *Dialog-Based Argumentation System* (D-BAS) developed by Krauthoff et al. (2017). Here, users are not shown the complete list of available

[1] `https://www.kialo.com/`, accessed 9 Nov 2020
[2] `https://reasonscore.com/`, accessed 9 Nov 2020

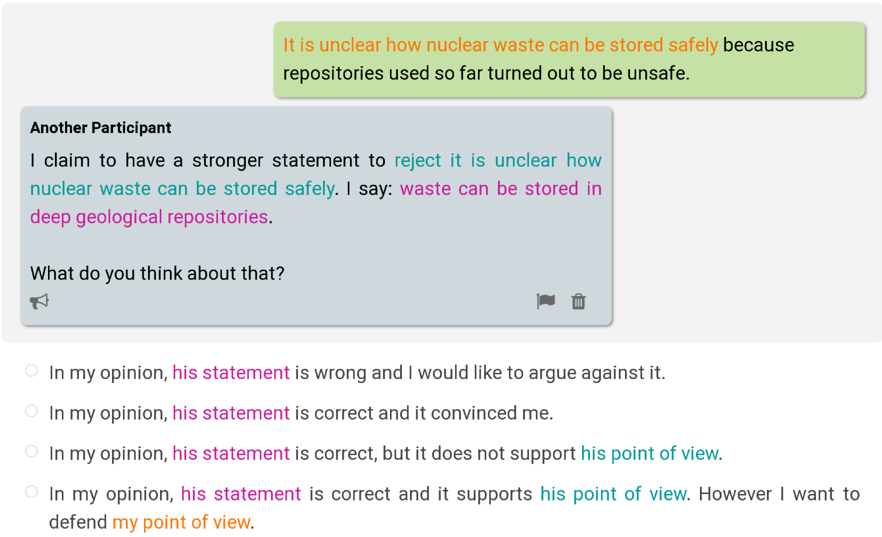


Figure 2.3: An excerpt from a dialog with D-BAS.

arguments, but a conversation with a chat partner is simulated, as shown in Figure 2.3. The virtual chat partner takes arguments from a corpus of arguments provided by other user, and asks the user to choose existing or add new arguments to defend their own views. The UI also allows undercut attacks, which are presented user-friendly, suited for untrained people, so that users have not to be familiar with the terms from argumentation theory.

Different applications have been built based on D-BAS: *discuss* (Meter et al., 2017) allows embedding D-BAS into arbitrary websites. The software *decide* (Ebbinghaus and Mauve, 2020) has been built for making collaborative decisions with many participants after exchanging arguments. Last but not least, the *Extensible Discussion Entity Network* (EDEN) by Meter et al. (2018b) enables the use of a distributed, shared argumentation graph within different front-end applications.

Instead of dedicated websites, there are also chat bots built on argumentation tools, which can make the argumentation experience feel more natural by picking people up at established places, and which enable different kinds of automatic interventions from the virtual chat partner. The bot Jebediah by Meter et al. (2018a) can be used to chat with D-BAS within social networks like Facebook, and Correnz (2020) built a D-BAS integration for the messenger Telegram. Inspired by D-BAS, Chalaguine and Hunter (2020) developed a persuasive chat bot. Their bot tries to understand a user's concerns and looks for suitable arguments in its argumentation graph. The chat system PEOPLES by Ittermann and Plüss (2020) aims at intervening if discussions become too polarized, hence, hopefully, leading to a healthier discussion atmosphere.

To sum up, different applications which create and exploit the structure of argumentation graphs have been developed by different researchers. Many applications intent to provide a structured database of arguments, but more sophisticated applications like persuasive bots have also arisen.

Chapter 3

Developing a Metric for Comparing Attitudes in Argumentations

Many of us have already taken part in discussions on the Internet and have read other people's arguments and stated own arguments. Sometimes, you might have asked yourself: How similar are my own attitudes compared to those other people? Who has attitudes which match my own attitudes most? Or which organizations share my own arguments, or have a completely different view?

To answer those questions, we have to calculate the distance between different attitudes in an argumentation, but there have not been any means to do a sensible comparison so far. Therefore, we have a look at how to calculate the (dis)similarity of lines of arguments in this chapter. Such a similarity measure has applications in building neighborhood-based recommender systems considering argumentation behavior, finding people with similar or different argumentation behavior, or comparing the opinions of political parties and voters.

We will first motivate and explain why dealing with this question is pertinent and how the relevant pieces of information can be represented mathematically. Thereafter, we expound on our proposal for a pseudometric for calculating the distance between two attitudes in an argumentation. We will then examine how to make sure that such a distance function yields results matching human intuition. Lastly, we compare different distance functions regarding their intuitiveness and study how much our proposed pseudometric is in line with intuition. As we will see, there are different suitable "best" functions, depending on the application context.

3.1 Capturing Personal Attitudes in Weighted Argumentation Graphs

In this section, we delve into our model of weighted argumentation graphs, which aims at representing individual attitudes stated within an argumentation. After a general introduction, we discuss different aspects and design decisions of the model and have look at similar models proposed in earlier research. A more formal definition of the model will be given in Subsection 3.2.2.

3.1.1 An Introduction to Weighted Argumentation Graphs

Within an argumentation, a person's (let us call they Alice) attitude on the topic discussed could be—in a condensed form—typically expressed like this:

> I am against building more nuclear power plants; I am very sure of it. My main reason is that there are no safe disposal sites. I am also quite sure that nuclear power plants are unsafe. On the other hand, I agree that nuclear power is sustainable, but I consider this counterargument unimportant in comparison.
>
> Also, I am reasonable sure that planting genetically modified crops should be allowed. This is less important to me than the question of building more nuclear power plants.

From this example, we can deduce which parts have to be considered when capturing Alice's expression of her view:

1. She states whether she accepts or rejects statements.
2. She says how sure she is when accepting or rejecting a statement.
3. Her arguments for/against certain statements are mentioned.
4. The relative importance of arguments is expounded on.
5. She indicates the relative importance of positions.

Hence, if we want to compare Alice's attitudes with someone else's attitudes, we have to find a suitable model capturing all these pieces of information. Take, for example, the relative importance of positions: If I am strongly in favor of nature conservation, then my attitude is probably closer to that of a green political party than the attitude of a liberal party, even if both parties have the same opinion on the position, because the topic has a higher priority for a green party.

We now explain our new model of weighted argumentation graphs (Brenneis et al., 2020), which is based on the IBIS model (Kunz and Rittel, 1970). To model Alice's attitude in the argumentation for comparison with other attitudes, we first create an argumentation graph containing all arguments mentioned in that argumentation by any participant. For our example, this could result in the graph G in Figure 3.1.

Every participant in the argumentation can have a personal view on that graph which contains all statements and arguments communicated by that person. So Alice's graph G_A in Figure 3.1 only contains the nodes for statements she mentioned above.

Furthermore, we add ratings to each statement node with a value in $[-0.5, 0.5]$, which express how sure the person is that the statement is true or not. A positive value indicates that a

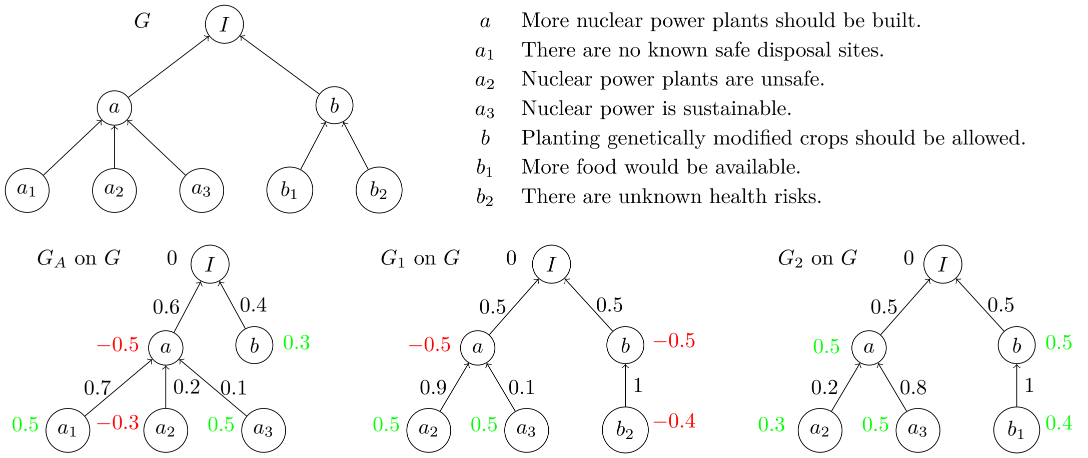


Figure 3.1: Examples of an argumentation graph G and several weighted argumentation graphs on G (G_A, G_1, G_2).

person accepts a statement, a negative value expresses disagreement. A value of 0 is treated as default value and indicates an unknown opinion or neutrality.

The relative importance of arguments is modeled by weights on the argument edges. Arguments with a higher weight are considered more important in comparison to other arguments with the same conclusion. An argument not mentioned has an implicit weight of 0. The weights of arguments with the same conclusion must sum up to 1 or 0.

The relative importance of positions is also modeled as "argument" weight, although edges like (a, I) are no real arguments. An intuitive interpretation of this choice is reinterpreting the issue node I as "personal well-being," which then allows to see (a, I) as an argument: "My personal-well being improves, because no more nuclear power plants are built."

When we refer to the sum of personal argument weights and statement ratings, we use the term attitude in this chapter. Other persons or organizations (e.g. political parties) can have different attitudes towards the statements and arguments in G. The graphs G_1 and G_2 may represent the attitudes of Party 1 and Party 2, respectively. As all those graphs are based on G, we say that G_A, G_1, and G_2 are a weighted argumentation graph on G.

3.1.2 Discussion on Aspects of the Model

Let us now regard why certain aspects of the model were chosen that way. For many decisions, we had in mind that the model has to work in an application context, i.e. the weights and ratings have to be collected somehow from a user.

Our model is based on IBIS because it is a simple and intuitive model for everyday argumentation around different issues. It has been shown that argumentation systems built on this model can be handled by untrained users with a suitable UI, for example with D-BAS, which was used in different field experiments (Ebbinghaus, 2019; Krauthoff et al., 2017).

Moreover, this model uses statements as atomic building blocks. This choice allows us to attach individual persons' opinion ratings for statements to a node and weights for arguments on the edges. As there is a dedicated root node I, we also capture "how far away" a statement is from the central positions, i.e. how long the paths from statement nodes to I are. Therefore, we are able to give deeper nodes a lower weight when comparing graphs later on.

Although everyday usability should not limit the complexity of the model, it is important to keep in mind that every number in the model has to be collected somehow from the users. For example, it would be impractical if the model and the metric required opinions on every statement in the argumentation graph since users of argumentation systems usually only give opinions on a small subset of statements. In addition, keeping a model as simple and as easy to understand as possible is a sensible goal in itself.

The range of ratings is limited to the interval $[-0.5, 0.5]$ for two reasons: First, the diameter 1 will later turn out to be handy for having distances which are always $[0, 1]$, and thus also always finite. This effect could also be achieved with any other interval by rescaling values within the metric, but its formula will be simpler by choosing the fitting interval beforehand.

In addition, there is a practical reason for specifying a finite interval: If a user had to assess how sure they are about their opinion, it would be easier to give values on a limited scale like a Likert scale instead of specifying a value from an unspecified range.

For similar reasons, the sum of the argument weights leading to a common conclusion must be 1. Furthermore, for arguments, we are only interested in the relative importance to the conclusion. Having an absolute weight for arguments feels unintuitive since one usually says that certain arguments are more important than others, not that an argument is "90 % important."

Whether an argument is an attacking or supporting argument is not included in our definition of weighted argumentation graphs. The plain reason is that we will not need this information later on in our pseudometric. What is more, whether supporting or attacking arguments are agreed to should have an influence on the user's rating for the statement. Therefore, this aspect is implicitly captured by the ratings. Nevertheless, a complete representation of an argumentation would have to capture whether an argument is attacking or defending.

Another aspect is that we cannot model undercut attacks, i.e. attacks with an argument (edge) as conclusion. We do not think that undercuts play an important role in every application context, hence we left them out in our model and metric proposal. As point of fact, undercuts were rarely used in field experiments with D-BAS (Ebbinghaus, 2019; Krauthoff et al., 2017).

Furthermore, an application could treat undercuts like undermines. Let us consider our example from Section 2.1 again: "One's own opinion should not be based on who made a proposal." is an undercut attack to "We should build more nuclear power plants because Party X suggested doing so." This attack could also be presented as targeting "Party X suggested doing so." as long as the undercut relation is maintained in the backing database. Such an approach is taken in the interface of *decide* by Ebbinghaus and Mauve (2020).

Last of all, it makes sense to require consistency of argument weights and agreement ratings for statements at first sight. For instance, a statement should be accepted if there are only supportive arguments. In theory, ratings might even be calculated from the ratings of the premises and the weights of the arguments, possibly in a way similar to Selinger (2014). Real world discussions, however, can contain inconsistent attitudes, e.g. when falling for fallacies of inconsistency (Damer, 2008). Furthermore, it is hard to collect all considered arguments from a user. As we want to be able to model non-ideal real-world argumentations, we impose no consistency constraints in our model.

3.1.3 Related Work

Different definitions of argumentation graphs with some kind of weights have been made before, basing upon Dung's definition of argumentation frameworks (Dung, 1995), but they differ in several aspects from our model which made them unsuitable for our application. Remember that in Dung's definition, arguments (not statements) are the nodes and attacks the edges.

Amgoud et al. (2017) and Gordon and Walton (2016) extended Dung's model with a weight function assigning every argument a value between 0 and 1; note that Amgoud et al. (2017) also call their model *weighted argumentation graph*, which should not be confused with our definition. In a work by Hunter (2013), each argument in an argumentation framework was associated with a probability of being believed to be true. Martínez et al. (2008) defined an abstract argumentation framework with a binary relation over the attacks which specifies the "strength" of the attack by assigning a value between 1 and n. Dunne et al. (2011) suggested assigning a positive real valued weight to argument nodes and leaved the interpretation of this weight open to the application context. *Bipolar Weighted Argumentation Frameworks* (BWAF) were defined by Pazienza et al. (2017a) to capture attacks and defenses together with a strength from a bounded interval. The definition and motivation of epistemic graphs by Hunter et al. (2020) is similar to our thought of having personal views on an argumentation.

All those definitions include, however, only one kind of numeric value, whereas we need to differentiate between ratings for statements and weights of arguments. Therefore, a new kind of data structure was necessary.

After we had published our work, Ferilli (2020) presented the idea of the *Generalized Argumentation Framework* (GAF) which also has weights for both, nodes and vertices, and can incorporate the confidence of persons into arguments. They also introduced a matrix notation similar to our notation we will use in Subsection 3.2.1. Additionally, they modeled attacks as well as defenses. Their work is based on their earlier proposals of BWAFs and Trust-affected BWAFs (Pazienza et al., 2017b). In contrast to our work, they do not allow negative weights for nodes, but they do allow negative weights for edges. This incompatibility could be worked around by using their user trust function, such that the following thoughts about our weighted argumentation graphs could be mapped to GAFs.

3.2 Towards a Metric on Weighted Argumentation Graphs

Now we move on to the question of how we can actually calculate the distance between two attitudes which are expressed as weighted argumentation graphs. This means that we are interested to know how close the *views* of two people or agents are concerning the positions discussed; we do *not* want to compare their style of argumentation, e.g. regarding consistency, stylistics devices, or number of arguments used. Therefore, we will only compare weighted argumentation graphs which have the same underlying unweighted graph.

We now introduce a pseudometric for this comparison task. First, we explain the intuition behind our pseudometric in a thorough example. Afterwards, we formally define our pseudometric by considering our publication (Brenneis et al., 2020) about that pseudometric. We also explain which properties we expect of a suitable metric and prove that our proposal fulfills them. Finally, we outline some further limitations which should be considered when working with such a distance function.

3.2.1 Introducing Our Pseudometric

There are already different means for comparing (weighted) graphs or trees outside the field of argumentation, but they fail to deal with particular features of argumentations. For example, the common tree edit distance (Bille, 2005) is not suitable for our application, as the depth of nodes plays no role, but we feel that opinions on "deeper" statements should have less influence than opinions on positions. Moreover, we do not need to measure structural similarity, since the structure of the underlying unweighted argumentation graphs of the weighted argumentation graphs being compared is always the same. Hence, the metric for directed weighted graphs by Xu et al. (2013) is also unsuitable.

Let us continue our example from Figure 3.1, where we considered the attitudes of Alice and two political parties. If a VAA like the German Wahl-O-Mat had to compare the attitudes of Alice and the parties, the overlap of Alice with the parties would be 50% in both cases (if we ignore weighting of positions) since the opinions on exactly one of two positions is the same and arguments are not taken into account. So ignoring the additional information present in a weighted argumentation graph is no option.

When considering argument weights and statement ratings, we want to preserve some intuitive properties, e.g. opinions on arguments should carry less weight than opinions on positions. We will later call such properties, which we consider intuitive and should be fulfilled by any metric comparing attitudes in argumentations, *desiderata*. Our list of desiderata has been used to check metric candidates for sensibility during research.

Our proposed pseudometric calculates the distance between Alice (G_A) and Party 1 (G_1) as follows: For every statement s, we multiply its rating with the product of the weights on the path from s to I. The absolute differences for those values in G_A and G_1 are added, weighted by α raised to the depth of s to reduce the weight of nodes deeper in the graph. Hence, a lower α emphasizes the opinions of the positions, a higher value emphasizes opinions deeper in the graph. Thus, α is similar to the damping factor in the PageRank algorithm (Page et al., 1999). The result is normalized with $1 - \alpha$ so that it is always in $[0, 1]$. We will prove this fact and provide a formal definition of the pseudometric later on.

If we choose $\alpha = \frac{1}{2}$, we get the following distance between Alice and Party 1:

$$
\begin{aligned}
d_G(G_A, G_1) = \left(1 - \frac{1}{2}\right) \cdot \Bigg[&\left(\frac{1}{2}\right)^0 \cdot \underbrace{|0 - 0|}_{I:\ 0} && (3.1)\\
+ &\left(\frac{1}{2}\right)^1 \cdot \underbrace{|-0.5 \cdot 0.6 - (-0.5) \cdot 0.5|}_{a:\ 0.05} && (3.2)\\
+ &\left(\frac{1}{2}\right)^2 \cdot \underbrace{|0.5 \cdot 0.7 \cdot 0.6 - 0 \cdot 0 \cdot 0.5|}_{a_1:\ 0.21} && (3.3)\\
+ &\left(\frac{1}{2}\right)^2 \cdot \underbrace{|-0.3 \cdot 0.2 \cdot 0.6 - 0.5 \cdot 0.9 \cdot 0.5|}_{a_2:\ 0.261} && (3.4)\\
+ &\left(\frac{1}{2}\right)^2 \cdot \underbrace{|0.5 \cdot 0.1 \cdot 0.6 - 0.5 \cdot 0.1 \cdot 0.5|}_{a_3:\ 0.005} && (3.5)\\
+ &\left(\frac{1}{2}\right)^1 \cdot \underbrace{|0.3 \cdot 0.4 - (-0.5) \cdot 0.5|}_{b:\ 0.37} && (3.6)\\
+ &\left(\frac{1}{2}\right)^2 \cdot \underbrace{|0 \cdot 0 \cdot 0.4 - 0 \cdot 0 \cdot 0.5|}_{b_1:\ 0} && (3.7)\\
+ &\left(\frac{1}{2}\right)^2 \cdot \underbrace{|0 \cdot 0 \cdot 0.4 - (-0.4) \cdot 1 \cdot 0.5|}_{b_2:\ 0.2} \Bigg] && (3.8)\\
= 0.1895 & && (3.9)
\end{aligned}
$$

This calculation can be implemented recursively using depth-first search for argumentation *trees*, which have no cycles. Whenever going a step deeper, a factor α is added.

The same calculation can also be expressed using a matrix of argument weights and a vector of statement ratings. Let us define the following weight matrices representing the argument weights of G_A and G_1:

$$w_A := \begin{array}{c} \\ I \\ a \\ a_1 \\ a_2 \\ a_3 \\ b \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cccccccc} I & a & a_1 & a_2 & a_3 & b & b_1 & b_2 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{3.10}$$

$$w_1 := \begin{array}{c} \\ I \\ a \\ a_1 \\ a_2 \\ a_3 \\ b \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cccccccc} I & a & a_1 & a_2 & a_3 & b & b_1 & b_2 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.0 & 0 & 0 \end{pmatrix} \end{array} \tag{3.11}$$

The rows represent the premises, the columns conclusions, i.e. the value 0.6 in w_A is the weight for the argument (a, I) in G_A, where I is the conclusion. Each summand in the following calculation summarizes the contribution of paths with the same length to the overall sum, i.e.

the first summand is the contribution of all nodes which have a path of length 1 to I, which are the statements a (contribution 0.05) and b (0.37) in our example:

$$d_G(G_A, G_1) = \left(1 - \frac{1}{2}\right) \cdot \left[\left(\frac{1}{2}\right)^1 \left\| w_A[:,1] \odot \begin{pmatrix} 0 \\ -0.5 \\ 0.5 \\ -0.3 \\ 0.5 \\ 0.3 \\ 0 \\ 0 \end{pmatrix} - w_1[:,1] \odot \begin{pmatrix} 0 \\ -0.5 \\ 0 \\ 0.5 \\ 0.5 \\ -0.5 \\ 0 \\ -0.4 \end{pmatrix} \right\|_1 \right. \tag{3.12}$$

$$\left. + \left(\frac{1}{2}\right)^2 \left\| w_A^2[:,1] \odot \begin{pmatrix} 0 \\ -0.5 \\ 0.5 \\ -0.3 \\ 0.5 \\ 0.3 \\ 0 \\ 0 \end{pmatrix} - w_1^2[:,1] \odot \begin{pmatrix} 0 \\ -0.5 \\ 0 \\ 0.5 \\ 0.5 \\ -0.5 \\ 0 \\ -0.4 \end{pmatrix} \right\|_1 \right] \tag{3.13}$$

$$= \left(1 - \frac{1}{2}\right) \cdot \left[\left(\frac{1}{2}\right)^1 \left\| \begin{pmatrix} 0 \\ -0.05 \\ 0 \\ 0 \\ 0 \\ 0.37 \\ 0 \\ 0 \end{pmatrix} \right\|_1 + \left(\frac{1}{2}\right)^2 \left\| \begin{pmatrix} 0 \\ 0 \\ 0.21 \\ -0.261 \\ 0.005 \\ 0 \\ 0 \\ 0.2 \end{pmatrix} \right\|_1 \right] \tag{3.14}$$

$$= 0.1895 \tag{3.15}$$

The values in the vectors in (3.14) are the same values as in the absolute values in terms (3.1) et seq.

This matrix representation has the advantage that it works more generally on argumentation graphs, where statements can be reused and cycles can be built. In practice, when calculating the distance between graphs containing at least one cycle, the sum can be truncated once a suitable precision has been reached. The factor α assures convergence of the sum, as we will show in a second.

For the distance between Alice's argumentation graph and the argumentation graph of Party 2, we get $d_G(G_A, G_2) = 0.25075$. So Party 1 and Alice are closer to each other than Party 2 and Alice because the opinions on nuclear power, which is more important for Alice, are the same, although the arguments of Party 1 are different. If the arguments used were more similar, the distance would be even smaller, which matches our intuition.

To get a better feeling of the absolute values our pseudometric produces, we can have a look at the maximum possible value. For every depth D, we get, at maximum, a summand of $(1-\alpha)\cdot 1+\alpha\cdot d_{D+1}$, where d_{D+1} is the contribution of deeper depths, and the factor 1 is the maximum possible absolute difference between two (weighted) opinion ratings. For the matrix version for graphs, remember that the power of a matrix with non-negative entries and column sums less than or equal to 1 (i.e. a substochastic matrix) produces a matrix with the same property. Because the geometric series converges, we can calculate the upper bound

$$\sum_{D=1}^{\infty}\alpha^D(1-\alpha)\cdot 1=\frac{\alpha}{1-\alpha}(1-\alpha) \tag{3.16}$$

$$=\alpha. \tag{3.17}$$

As our example is a tree (i.e. no cycles and finite depth), the recursive definition of the pseudometric yields no more positive values at depth $D>2$, resulting in a stricter upper bound:

$$\sum_{D=1}^{n}\alpha^D(1-\alpha)\cdot 1=\frac{\alpha(1-\alpha^n)}{1-\alpha}\cdot(1-\alpha) \tag{3.18}$$

$$=\alpha\cdot(1-\alpha^n) \tag{3.19}$$

$$\overset{n=2}{=}\frac{1}{2}\cdot\left(1-\frac{1}{2^2}\right) \tag{3.20}$$

$$=0.375 \tag{3.21}$$

So if we want to have a percentage value for the dissimilarity of the attitudes, we could say that Alice and Party 1 have a dissimilarity of $\frac{0.1895}{0.375}\approx 51\%$, and Alice and Party 2 a dissimilarity of $\frac{0.25075}{0.375}\approx 67\%$. For many applications, though, the absolute values are not as important as the resulting sort order.

The calculations of the pseudometric can also be interpreted as follows: For argumentation trees, the distance function can be seen as an embedding of the argumentation tree in an n-dimensional Euclidean space, where n is the number of statements in G, and the value of each component is equal to the statement's s rating multiplied with the weights of the arguments leading to the root, and then calculating the L_1 norm. This interpretation makes clearer that some basic properties of a metric (e.g. the triangle inequality) are fulfilled.

Another interpretation is a random walk: The matrix w gives the probability of following a certain path, α is the probability of moving to a child node. w^2 gives the probability after 2 steps etc.; the probability of going exactly i steps is $(1-\alpha)\cdot\alpha^i$, which sums up to 1 for an infinite number of steps. This interpretation allows applying known stochastic techniques in future work.

3.2.2 Formal Introduction of Our Model, Pseudometric, and Desiderata

After this introduction by example, we now formally introduce our weighted argumentation graphs as well as our pseudometric. For this propose, we have a look at our peer-reviewed conference paper in which we presented them for the first time:

Markus Brenneis, Maike Behrendt, Stefan Harmeling and Martin Mauve.
"How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs"
In: *Proceedings of the 3rd International Workshop on Systems and Algorithms for Formal Argumentation*, pages 2–13, CEUR Workshop Proceedings Volume 2672.
Acceptance Rate: 90%

A verbatim copy of the published paper is included in this section. In the paper, we made the following key contributions:

1. introduction of the new model of weighted argumentation graphs, with weights for both statements and arguments
2. a pseudometric to calculate the distance between two weighted argumentation graphs
3. a set of intuitive desiderata a metric for weighted argumentation graphs should fulfill
4. proofs that our pseudometric fulfills those desiderata

Personal Contribution

The ideas for weighted argumentation graphs with two kinds of weights, constructing a recursive metric on argumentation graphs, and maintaining a list of desiderata to identify possible unsuited metrics were developed by the author of this thesis, Markus Brenneis. The concepts were discussed and improved on in meetings with Maike Behrendt, Stefan Harmeling, and Martin Mauve, where Stefan Harmeling provided a stochastic interpretation of the model. Markus Brenneis wrote the whole paper. Maike Behrendt, Stefan Harmeling, and Martin Mauve provided feedback on the drafts, where Stefan Harmeling focused on feedback regarding the presentation of proofs.

Importance and Impact on This Thesis

In this paper, we present the first pseudometric for measuring the distance between argumentation graphs which have both weights for statements and arguments. Having this pseudometric is prerequisite for the recommender system in our *deliberate* system, which we introduce in Section 4.1, and the calculation of voter–party similarity in our VAA presented in Section 4.2. Whether the desiderata proposed in this paper look not only useful for scientists, but are intuitive for people without background knowledge in argumentation theory, was evaluated in later publications, which we present in Sections 3.3 and 3.4.

How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs

Markus BRENNEIS [a,1], Maike BEHRENDT [a] Stefan HARMELING [a] and Martin MAUVE [a]
[a] *Department of Computer Science, University of Düsseldorf, Germany*

Abstract. When exchanging arguments with other people, it is interesting to know who of the others has the most similar opinion to oneself. In this paper, we suggest using weighted argumentation graphs that can model the relative importance of arguments and certainty of statements. We present a pseudometric to calculate the distance between two weighted argumentation graphs, which is useful for applications like recommender systems, consensus building, and finding representatives. We propose a list of desiderata which should be fulfilled by a metric for those applications and prove that our pseudometric fulfills these desiderata.

Keywords. argumentation graphs, online argumentation, metric

1. Introduction

In real-world discussions, people exchange arguments on a dedicated issue, such as improving the course of study [11], the distribution of funds [7], or which party to vote for at the next general election. In all those cases, participants discuss positions like "there should be a universal basic income" or "special math courses should be introduced", state their pro and contra arguments and attack other people's arguments.

Each individual participant of an argumentation has a personal view on the arguments and their relative importance: Users can decide for themselves which arguments they consider more convincing, thus which arguments they agree to and how much they agree with a statement. They may consider some positions more important than others.

Based on those individual views, there are useful applications for measuring the similarity or distance between two users: Clustering can be used to find representatives for a group of people with similar argumentation behavior or for finding a consensus. Another application is opinion polling, where one wants to find out why two persons or organizations come to different conclusions. What is more, collaborative filtering, which needs some definition of distance between users, can be used for pre-filtering arguments in applications like Kialo[2].

---

[1]Corresponding Author: Markus Brenneis, Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany; E-mail: markus.brenneis@uni-duesseldorf.de.

[2]https://kialo.com/

In this paper, we propose solutions to the two main challenges to achieving the goal of comparing the argumentations of two users: We define weighted argumentation graphs which are a suitable representation of argumentation covering the mentioned aspects, including importance of arguments and agreement with statements. Secondly, we suggest a pseudometric for calculating the distance between two weighted argumentation graphs, which considers the specific structure of argumentation graphs (e.g., opinions deeper in a graph are less important). We contribute a list of useful desiderata for a metric which compares argumentations, and prove that our pseudometric fulfills those properties.

In the following chapter, we present our definition of weighted argumentation graphs. The third chapter introduces our pseudometric and desiderta for a useful metric. Finally, we discuss some limitations of our pseudometric and take a look at related work.

2. Definition of Weighted Argumentation Graphs

To be able to determine the similarity of real-world argumentations, there has to be a suitable representation of them. This representation should be able to capture all aspects mentioned in the introduction, and it should be as simple as possible. Therefore, the following definition is based on the IBIS model [12], which has been successfully tested with users without background in argumentation theory using our D-BAS system [11].

For the application purposes described in the introduction, the model should be able to represent the known opinions and arguments of a person as close as possible. Thus, we use *statements*, not arguments, as atomic elements in our definition, which then can be composed to arguments which can support or attack another statement. Note, though, that this definition can be translated to classical abstract argumentation frameworks based on Dung's definition [5], e.g. using DABASCO [14].

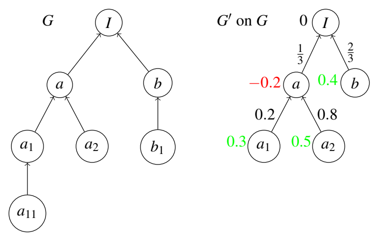
Let S be the (finite) set of all statements, with the special statement $I \in S$. The set $A \subseteq S \setminus \{I\} \times S$ is a set of arguments. For an argument $a = (s_1, s_2) \in A$, s_1 is called premise, s_2 conclusion of a. Let $s \in S$, then $a_{\to s} := \{(t,u) \in A \mid u = s\}$ is the set of arguments with conclusion s.

Note that I is excluded to be a premise since it is the *issue* in the IBIS model. We refer to I as "personal well-being", which allows us to interpret an edge like (b,I) in Figure 1 as "My personal well-being improves, because more wind power plants will be built." The premises of arguments with conclusion I are called positions. Positions are actionable items like "A wall between Austria and Germany should be built," and play an important role in real-world argumentation, e.g. decision-making problems [7].

Definition 1 (Argumentation Graph)**.** *An* argumentation graph *is a directed, weakly connected graph $G = (S,A)$, $A \subseteq S \setminus \{I\} \times S$, where the statements are nodes and the arguments edges, and there is exactly one $I \in S$ which has no outgoing edges.*

Note that this model does not include different relations for attack and support, as known from bipolar argumentation frameworks. Whether an argument is supportive are not, is up to a person's interpretation of the natural language representation of the arguments. The purpose of the model is solely to capture the hierarchy of statement, which we later need for our metric; bipolarity would add unnecessary complexity in this paper.

Every person can have a personal view with personal attitudes on a common argumentation graph G, as depicted in G' in Figure 1. To get an intuition for our next defini-



a	A wall between Austria and Germany should be built.
a_1	The wall stops illegal immigration of cows.
a_{11}	Cows can come via Switzerland.
a_2	The wall is expensive.
b	More wind power plants should be built.
b_1	Wind power is a renewable energy source.

Figure 1. Example for an argumentation graph G and a weighted argumentation graph G' on G with positions a and b and concrete examples for each statement. Edges with weight 0 and nodes with rating 0 are not drawn. Statement ratings are next to nodes, values for relative argument importance next to edges.

tion, let us have a look at what we can deduce about Alice's attitudes from her graph G': Alice strongly accepts position b (rating .4) and is slightly against a (rating $-.2$). She accepts the statement a_2 more than statement a_1 (rating $.5 >$ rating .3). The counterargument (a_2, a) "No wall should be built, because a wall is expensive." is far more important for her than the argument (a_1, a) (relative importance $.8 > .2$).

Furthermore, it makes sense to sort the positions: She considers building more wind power plants (b) more important than building a wall (a, relative importance $\frac{2}{3} > \frac{1}{3}$). So when comparing her attitudes with someone else's attitudes, she would consider a contrary opinion on b more severe than a different opinion on a. For ordinary statements, which are not positions, having an importance does not make sense: One cannot say that "The wall stops illegal immigration of cows" is twice as important as "Cows can come via Switzerland" (important regarding what?); one can only say that the arguments regarding building a wall which are built by those statements are of differing importance.

We will use real numbers to represent those weights and ratings.

Definition 2 (Weighted Argumentation Graph)**.** *Let $G = (S, A)$ be an argumentation graph. A* weighted argumentation graph *G' on G is a quadruple (S, A, r, w) with functions r and w. $r\colon S \to [-0.5, 0.5]$ assigns an agreement score (rating) to every statement, where negative values mean disagreement,* 0 *no opinion/don't care, and positive values agreement. $w\colon A \to [0,1]$ assigns an importance weight to each argument. The value indicates the importance of that argument relative to other arguments with the same conclusion. The value* 0 *means that the argument is not used (i.e. has no relevance), and* 1 *means that the argument is the only relevant argument for the conclusion. The following conditions must hold:*

$$\forall s \in S \sum_{a \in A_{\to s}} w(a) \in \{0, 1\} \tag{1}$$

$$r(I) = 0 \tag{2}$$

Formula 1 means that the sum of weights of arguments with the same conclusion is 1 if there is an argument with positive weight (cf. (a_1, a) and (a_2, a) in Figure 1); the sum is 0 iff no argument for a common conclusion has a weight (cf. $w(b_1, b) = 0$ in Figure 1). This assures that w represents *relative*, not absolute importance. To simplify notation, we write $w(\cdot, \cdot)$ instead of $w((\cdot, \cdot))$. If the underlying argumentation graph G happens to be a directed tree, we call G' a weighted argumentation tree.

w and r can be represented as matrix or vector, respectively, where undefined values are set to the default value 0. For the example in Figure 1, one gets:

$$w = \begin{pmatrix} 0 & 0 & 0\,0\,0\,0\,0 \\ 0.33 & 0 & 0\,0\,0\,0\,0 \\ 0.67 & 0 & 0\,0\,0\,0\,0 \\ 0 & 0.2 & 0\,0\,0\,0\,0 \\ 0 & 0.8 & 0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0 \end{pmatrix}, \qquad r = \begin{pmatrix} 0 \\ -0.2 \\ 0.4 \\ 0.3 \\ 0.5 \\ 0 \\ 0 \end{pmatrix} \tag{3}$$

The entry in row i column j of w is the weight of the argument with premise i and conclusion j. The first column and first row must refer to I as premise or conclusion, respectively. Because of Formula 1, the column sum is always 0 or 1.

If we draw or talk about a weighted argumentation graph, "non-existing" edges a are edges with $w(a) = 0$ (argument with no importance), and "non-existing" nodes s are nodes with $r(s) = 0$ (neutral statement). An example is G' shown in Figure 1.

The importance of a position p is represented as weight of the "argument" (p,I) leading to the "personal well-being" I. An application could obtain weights and ratings from a user, for example, by asking them to mark statements which are considered more important, or sorting arguments by relevance, which we do in our *deliberate* system [4].

3. Proposal of a Pseudometric for Weighted Argumentation Graphs

We now propose a pseudometric for calculating a distance between two weighted argumentation graphs, and prove several properties we expect of a function which compares two argumentations. The goal of the metric is to indicated how close the opinions and used arguments of two persons are, considering graph structure and individual assessments of importance; we do *not* want to compare argumentations on abstract levels like consistency, number of arguments used, or if other person's arguments are countered.

3.1. The Pseudometric

We define a distance measure of two weighted argumentation graphs $G_1 = (S,A,r_1,w_1)$, $G_2 = (S,A,r_2,w_2)$ on $G = (S,A)$ as:

$$d_G(G_1,G_2) = (1-\alpha)\sum_{i=1}^{\infty} \alpha^i \|w_1^i[:,1] \odot r_1 - w_2^i[:,1] \odot r_2\|_1 \tag{4}$$

where $\alpha \in (0,1)$ determines the influence of opinions deeper in the graph: A lower α emphasizes opinions on statements $r(s)$, a higher value the similarity of the argumentation underneath a statement s. $w[:,1]$ denotes the first column of the weight matrix w, α^i is the i-th power of the scalar α, w^i the i-th power of the square matrix w, and $\odot$ the Hadamard (entrywise) product. The i-th summand calculates the contribution of the paths with length i ending at I. We drop the index G if the underlying argumentation graph is clear from the context.

The intuition behind this distance measure becomes clearer when rephrasing it for the special case of argumentation *trees* (which have no cycles or re-used statements, thus unique paths from each statement to I). In case of argumentation trees T_1 and T_2 on T, the distance d_G is equivalent to:

$$d_T(T_1,T_2) = (1-\alpha)\sum_{s\in S}\alpha^{\mathtt{depth}(s)}\left|r_1(s)\prod_{a\in\rho_{I\to s}}w_1(a) - r_2(s)\prod_{a\in\rho_{I\to s}}w_2(a)\right| \tag{5}$$

$\rho_{s_1\to s_2}$ is the sequence of all arguments (edges) on the path from s_1 to $s_2 \in S$ (where s_2 is deeper in the tree). $\mathtt{depth}(s) = |\rho_{I\to s}|$ is the length of the path from I to s, i.e. the number of arguments; $\mathtt{depth}(I) = 0$.

The terms in the absolute value measure the similarity of the opinions of a statement s as difference of their ratings, scaled with the product of the "importances" of the arguments leading to s. Hence, statements which are deeper in the argumentation tree get a smaller weight, and the overall relevance of an argumentation branch is limited to its importance.

To see how the calculation works and that the results match intuition, let us calculate the distance between the graphs G' (Figure 1), T_2, and T_3 (Figure 2) for $\alpha = 0.5$. We can expect that T_2 is closer to T_3 than to G', because the opinions on the statements a and b match and only the weights are different. The results confirm the expectation:

$$\begin{aligned}d(T_2,G') =&(1-\alpha)\left(\alpha^1\left|0.5\cdot 0.6-(-0.2)\cdot\frac{1}{3}\right|+\alpha^1\left|0.5\cdot 0.4-0.4\cdot\frac{2}{3}\right|\right.\\ &\left.+\alpha^2\left|0\cdot 0\cdot 0.6-0.3\cdot 0.2\cdot\frac{1}{3}\right|+\alpha^2\left|0\cdot 0\cdot 0.6-0.5\cdot 0.8\cdot\frac{1}{3}\right|\right)=0.194\end{aligned} \tag{6}$$

$$d(T_2,T_3) =(1-\alpha)(\alpha^1|0.5\cdot 0.6-0.5\cdot 0.3|+\alpha^1|0.5\cdot 0.4-0.5\cdot 0.7| = 0.075 \tag{7}$$

Note that the value of d_G and d_T is in $[0,1)$. If d is the depth of T, the maximum value of $d_T(T_1,T_2)$ is $\alpha(1-\alpha^d)$.[3] The maximum value of d_G is $\lim_{n\to\infty}(1-\alpha)\sum_{i=1}^n\alpha^i = \alpha$.

Theorem 1. *Let G be an argumentation graph. d_G is a pseudometric, i.e. has the following properties for all weighted argumentation graphs G_1, G_2, G_3 on G:*

(i) $d_G(G_1,G_1) = 0$

(ii) $d_G(G_1,G_2) = d_G(G_2,G_1)$ *(symmetry)*

(iii) $d_G(G_1,G_3) \leq d_G(G_1,G_2) + d_G(G_2,G_3)$ *(triangle inequality)*

Proof. d_G converges: $\sum_i \alpha^i$ is geometric series, which converges for $\alpha \in (0,1)$. The value of the L_1 norm cannot be greater than 1 because for each column sum σ of the w^i, we always have $0 \leq \sigma \leq 1$.

(i) holds because the same values are subtracted in the L_1 norm.

(ii) is given since the L_1 norm is symmetric.

As each summand fulfills the triangle inequality, d_G also fulfills (iii). □

However, d_G is not a metric, because $d_G(G_1,G_2)$ can be 0 even if G_1 is not equal to G_2: Consider G_1 where all statements are agreed to and every argument weight is 0, and G_2 where all statements have a rating of 0. Because weights and ratings are multiplied, the distance is 0 though the weighted argumentation graphs are different.

[3] Remember that $r_1(\mathtt{root}(T)) = r_2(\mathtt{root}(T))$ because $\mathtt{root}(T) = I$, and $r(I) := 0$.

3.2. Desiderata for a Metric for Weighted Argumentation Trees

Our pseudometric is only one of many possible metrics for the applications described. We present a list of intuitive desiderata which, as we think, should be fulfilled by any metric comparing two person's argumentations, and thus should be considered when constructing alternative metric proposals. It is, however, hard to capture intuitive properties in graphs which can contain circular references and re-used statements. Therefore, we focus on argumentation trees in this section. Field experiments like [11] have also shown that users seldom create cycles or re-use statements in different branches in real discussions.

After each desideratum, we prove that our pseudometric (Formula 5) fulfills it for weighted argumentation trees. We consider those properties important in many real-world application domains of a metric, albeit not everywhere, as pointed out in Section 5.

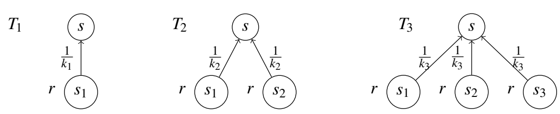
For each desideratum, we indicate why we think it is intuitive. Most desiderata are followed by a visual example making the choice of variable names clearer. Note that each tree in the examples is considered to be part of a bigger weighted argumentation tree, i.e. not all existing nodes and edges are drawn, and irrelevant statement ratings and argument weights are left out.

Desideratum 1 (Proportionally bigger overlap is better)**.** *Consider trees T_1, T_2, T_3, where T_2 is like T_1, but uses one additional argument for a statement s, and T_3 is like T_2, but uses one additional argument for s. Although T_2 and T_1 differ in only one argument, and T_3 and T_2 differ in only one argument, we expect $d(T_1, T_2) > d(T_2, T_3)$ because T_2 and T_3 have a greater overlap regarding the used arguments.*

More formally: For every statement s in a tree T which only has leaves $s_1, \dots, s_n$ ($n > 2$) as premises for s with $r(s_1) = \dots = r(s_n) \neq 0$ and $w(s_1, s) = \dots = w(s_n, s)$ and $\forall a \in \rho_{I \to s} : w(a) \neq 0$, consider the trees T_k, $n \geq k > 0$, which only contain $s_1, \dots, s_k$. Then, given $k_1 < k_2 < k_3$, we want to have $\frac{k_1}{k_2} < \frac{k_2}{k_3} \implies d(T_{k_1}, T_{k_2}) > d(T_{k_2}, T_{k_3})$, i.e. if the relative overlap of the number of arguments used is greater, the distance is smaller. Likewise, we demand $\frac{k_1}{k_2} > \frac{k_2}{k_3} \implies d(T_{k_1}, T_{k_2}) < d(T_{k_2}, T_{k_3})$, and $\frac{k_1}{k_2} = \frac{k_2}{k_3} \implies d(T_{k_1}, T_{k_2}) = d(T_{k_2}, T_{k_3})$.

We require $\forall a \in \rho_{I \to s} : w(a) \neq 0$ (i.e. no argument (edge) with weight 0 along the path from I to s), because if a user gives an argument a weight of 0, they say the premise is not related to the conclusion, thus not related to the topic of the discussion. This means that the user actually does not care about the opinions underneath that argument, which may be treated as if no opinion has been given.

In the following example, we have $k_1 = 1$, $k_2 = 2$, and $k_3 = 3$, thus $\frac{1}{2} < \frac{2}{3} \implies d_T(T_1, T_2) > d_T(T_2, T_3)$:



Proof. Only the argument weights for (s_i, s) (namely $\frac{1}{k_1}$, $\frac{1}{k_2}$, $\frac{1}{k_3}$, respectively, as used in (8) = (9)) are different and contribute to the sum, all other summands are zero. The common weight products and common values for r are summarized as $\overline{w_r}$ for readability and are factored out. For (10) > (11), remember that $\frac{k_1}{k_2} < \frac{k_2}{k_3}$.

$$d_T(T_{k_1}, T_{k_2}) = \overline{w_r}(1-\alpha) \sum_{s' \in \{s_1,\ldots,s_n\}} \alpha^{\mathtt{depth}(s')} |w_{k_1}(s',s) - w_{k_2}(s',s)| \tag{8}$$

$$= \overline{w_r}(1-\alpha) \sum_{s' \in \{s_1,\ldots,s_n\}} \alpha^{\mathtt{depth}(s')} \left(k_1 \cdot \left(\frac{1}{k_1} - \frac{1}{k_2} \right) + (k_2 - k_1) \cdot \frac{1}{k_2} \right) \tag{9}$$

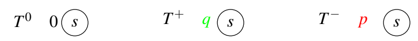
$$= \overline{w_r}(1-\alpha) \sum_{s' \in \{s_1,\ldots,s_n\}} \alpha^{\mathtt{depth}(s')} \left(2 - 2\frac{k_1}{k_2} \right) \tag{10}$$

$$> \overline{w_r}(1-\alpha) \sum_{s' \in \{s_1,\ldots,s_n\}} \alpha^{\mathtt{depth}(s')} \left(2 - 2\frac{k_2}{k_3} \right) = d_T(T_{k_2}, T_{k_3}) \tag{11}$$

The other cases are proven by replacing ">" with "<" or "=", respectively. □

Desideratum 2 (Contrary opinion is worse than no opinion)**.** *Consider trees T_1, T_2, T_3, where all trees are identical, but T_1 has no opinion on a statement s, T_2 a positive opinion on s and T_3 a negative opinion on s. As we definitely know that T_2 and T_3 disagree on s, we want to have $d(T_2,T_3) > d(T_1,T_2)$.*

Formally: For any statement s in a tree T with $\forall a \in \rho_{I \to s} : w(a) \neq 0$, let T^+ be like T but with $r^+(s) = q > 0$, T^- like T with $r^-(s) = p < 0$, and T^0 like T with $r^0(s) = 0$. Then $d(T^+,T^-) > d(T^+,T^0)$.



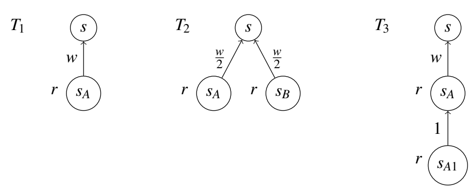
Proof. The only positive summand is the summand for s (which has rating 0, q or p for T^0, T^+, and T^-, respectively). The argument weights and the α term are common to all summands, can be factored out, and are summarized as $\overline{w_\alpha}$.

$$d_T(T^+,T^-) = \overline{w_\alpha} \cdot |q-p| = \overline{w_\alpha} \cdot (q+p) > \overline{w_\alpha} \cdot |q-0| = d_T(T^+,T^0) \tag{12}$$

□

Desideratum 3 (Deviation in deeper parts has less influence than deviation in higher parts)**.** *Consider the trees T_1,T_2,T_3, where T_1 has an argument (s_A,s) with no children and $\forall a \in \rho_{I \to s} : w(a) \neq 0$, T_2 is constructed from T_1 by adding a new statement s_B and argument (s_B,s) with $w_2(s_B,s) = w_2(s_A,s) = \frac{w_1(s_A,s)}{2}$, and T_3 is constructed from T_1 by adding a new statement s_{A1} and argument (s_{A1},s_A) with $w_3(s_{A1},s_A) = 1$. If $r(s_A) = r(s_B) = r(s_{A1}) \neq 0$, then we want to have $d(T_1,T_2) > d(T_1,T_3)$, because arguments deeper in the tree should have a smaller influence since we consider them less important for the overall opinion.*

We require $r(s_A) = r(s_B) = r(s_{A1}) \neq 0$ because adding a statement with "don't care" opinion should not actually change the distance. We want equality because this desideratum should cover only differences in the depth of the statements, not their rating.



Proof. The only differences are the summands including s, s_A and s_{A1}. Let $r := r(s_A) = r(s_B) = r(s_{A1}) \neq 0$ and $w := w_1(s_A, s)$. As before, we summarize common values for weights and α as $\overline{w_\alpha}$.

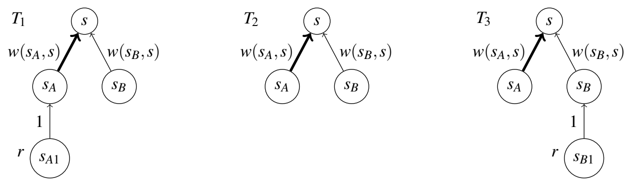
$$d_T(T_1, T_2) = \overline{w_\alpha} \cdot \left(\left| rw - r\frac{w}{2} \right| + \left| 0 - r\frac{w}{2} \right| \right) = \overline{w_\alpha} \cdot |rw| \tag{13}$$

$$> \overline{w_\alpha} \cdot \alpha |rw| = \overline{w_\alpha} \cdot (\alpha \cdot |r \cdot 0 \cdot w - r \cdot 1 \cdot w|) = d_T(T_1, T_3) \tag{14}$$

□

Desideratum 4 (Influence of deeper parts depends on weights in higher parts)**.** *Consider trees T_1, T_2, T_3, where all trees are identical, have statements s_A and s_B with conclusion s, and $w(s_A, s) > w(s_B, s)$, but T_1 has an additional argument for s_A and T_3 has an additional argument for s_B. Although the difference in both cases is only one argument, we expect $d(T_1, T_2) > d(T_3, T_2)$ because (s_A, s) has a larger weight.*

Formally: Let T_2 be a weighted argumentation tree with arguments $(s_A, s), (s_B, s)$ and $w(s_A, s) > w(s_B, s)$, no premises for s_A and s_B and $\forall a \in \rho_{I \to s} : w(a) \neq 0$. T_1 is constructed from T_2 by adding (s_{A1}, s_A) and T_3 from T_2 by adding (s_{B1}, s_B), each with a weight of 1 *and $r_1(s_{A1}) = r_3(s_{B1}) \neq 0$. Then $d_T(T_1, T_2) > d_T(T_3, T_2)$.*



Proof. Let $r := r_1(s_{A1}) = r_3(s_{B1}) \neq 0$. Only the summand which includes s_{A1} or s_{B1}, respectively, contributes a value greater than 0.

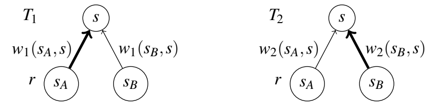
$$d_T(T_1, T_2) = \overline{w_\alpha} \cdot |w(s_A, s) \cdot r - 0| > \overline{w_\alpha} \cdot |0 - w(s_B, s) \cdot r| = d_T(T_2, T_3) \tag{15}$$

□

Desideratum 5 (Weights of arguments have influence even if they are the only difference)**.** *Consider trees T_1, T_2, where all trees are identical and have the arguments (s_A, s) and (s_B, s), but the weights are different: $w_1(s_A, s) \neq w_2(s_A, s)$ and $w_1(s_B, s) \neq w_2(s_B, s)$. We want to have $d(T_1, T_2) > 0$ if there exists a statement s' below s_A (or s_A itself) with $r_1(s') = r_2(s') \neq 0$ and $\forall a \in \rho_{I \to s'} : w(a) \neq 0$.*

We demand $r_1(s') = r_2(s') \neq 0$ because it makes sense if weights leading only to statements which are rated as "don't care" are ignored.

In this example, we have $s_A = s'$:



Proof. It is enough to show that there is at least one summand greater than 0.

$$|w_1(s_A,s)-w_2(s_A,s)|>0 \tag{16}$$

$$\implies \alpha^{\texttt{depth}(s')}(1-\alpha)|w_1(s_A,s)-w_2(s_A,s)|>0 \tag{17}$$

$$\implies \alpha^{\texttt{depth}(s')}(1-\alpha)r_1(s')\prod_{s_{A'}\in\rho_{s'\to I}\setminus(s_A,s)} w_1(s_{A'})|w_1(s_A,s)-w_2(s_A,s)|>0 \tag{18}$$

$$\implies \alpha^{\texttt{depth}(s')}(1-\alpha)\left|r_1(s')\prod_{s_{A'}\in\rho_{s'\to I}} w_1(s_{A'})-r_2(s')\prod_{s_{A'}\in\rho_{s'\to I}} w_2(s_{A'})\right|>0 \tag{19}$$

For (18) $\implies$ (19), remember that all weights in $\rho_{s'\to I}\setminus(s_A,s)$ and all ratings are the same for T_1 and T_2, e.g. $r_1(s')=r_2(s')$. □

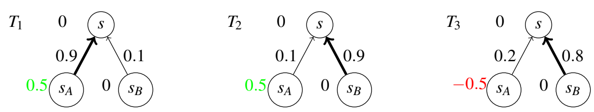
Desideratum 6 (Symmetry regarding negation of opinion)**.** *Let T_1, T_2 be any weighted argumentation trees, and T_3, T_4, respectively, the same trees, but the opinion for each statement is negated, i.e. $r_3(s)=-r_1(s)$ and $r_4(s)=-r_2(s)$ for all $s\in S$. We expect that a metric is symmetric regarding negation, i.e. $d(T_1,T_2)=d(T_3,T_4)$.*

Proof. This holds because $|r_1(s)-r_2(s)|=|(-r_3(s))-(-r_4(s))|=|r_3(s)-r_4(s)|$. □

Desideratum 7 (Trade-off between argument weights and agreement)**.** *Consider trees T_1,T_2,T_3 which are nearly identical and have leaf statements s_A and s_B with common conclusion s and $\forall a\in\rho_{I\to s}: w(a)\neq 0$. We have $r_1(s_A)=r_2(s_A)=0.5$, $r_3(s_A)=-0.5$, $r_1(s_B)=r_2(s_B)=r_3(s_B)=0$, and $w_1(s_A,s)>w_2(s_A,s)$. Furthermore, $w_2(s_B,s)=w_1(s_B,s)+w_1(s_A,s)-w_2(s_A,s)$, $w_3(s_B,s)=w_1(s_B,s)+w_1(s_A,s)-w_3(s_A,s)$, i.e. s_B is neutral and "collects" remaining weight such that the sum is 1.*

If $w_1(s_A,s)-w_2(s_A,s)>w_2(s_A,s)+w_3(s_A,s)$, although T_1 and T_2 have the same opinion on s_A, we want to have $d(T_1,T_2)>d(T_2,T_3)$, because both T_2 and T_3 do not care much about their (different) opinions on s_A. Likewise, if $w_1(s_A,s)-w_2(s_A,s)<w_2(s_A,s)+w_3(s_A,s)$, we expect $d(T_1,T_2)<d(T_2,T_3)$ because the weights w_1 and w_2 are closer to each other and give a greater weight for opposing opinions on s_A.

The following example trees depict the first case with concrete weight values. Because the different opinions on statement s_A are underneath an argument edge with small weight, we want to have $d(T_1,T_2)>d(T_2,T_3)$.



Proof. We proof the first case. For (20) = (21), remember that $w_1(s_A,s)>w_2(s_A,s)$.

$$\begin{aligned}
d(T_1,T_2) &= \overline{w_\alpha}\cdot|0.5\cdot w_1(s_A,s)-0.5\cdot w_2(s_A,s)| && (20)\\
&= \overline{w_\alpha}\cdot 0.5\cdot(w_1(s_A,s)-w_2(s_A,s)) && (21)\\
&> \overline{w_\alpha}\cdot 0.5\cdot(w_2(s_A,s)+w_3(s_A,s)) && (22)\\
&= \overline{w_\alpha}\cdot|0.5\cdot w_2(s_A,s)-(-0.5)\cdot w_3(s_A,s)| = d(T_2,T_3) && (23)
\end{aligned}$$

The other case follows by replacing ">" with "<". □

Desideratum 8 (Trade-off between statement ratings and agreement)**.** *Consider trees T_1, T_2, T_3 which are nearly identical and have a statement s and $\forall a \in \rho_{I \to s} : w(a) \neq 0$. We have $r_1(s) > r_2(s) > 0 > r_3(s)$ such that $|r_1(s) - r_2(s)| > |r_2(s) - r_3(s)|$. Although T_1 and T_2 have the same positive opinion on s, we want to have $d(T_1, T_2) > d(T_2, T_3)$, because both T_2 and T_3 have a weak opinion on s. Likewise, if $|r_1(s) - r_2(s)| < |r_2(s) - r_3(s)|$, we expect $d(T_1, T_2) < d(T_2, T_3)$ because the ratings $r_1(s)$ and $r_2(s)$ are closer to each other than $r_2(s)$ and $r_3(s)$.*

The first case, $d(T_1, T_2) > d(T_2, T_3)$, is shown in the following example:

T_1 0.4 s T_2 0.1 s T_3 −0.1 s

Proof. Only the summand for s contributes to the distance, all other summands are 0. Remember that all weights on the path to s are positive.

$$d(T_1, T_2) = \overline{w_\alpha}|r_1(s) - r_2(s)| > \overline{w_\alpha}|r_2(s) - r_3(s)| = d(T_2, T_3) \tag{24}$$

The other case follows by replacing ">" with "<". □

Desideratum 9 (Weights limit the influence of a path)**.** *Consider graphs T_1, T_2 which are nearly identical, have an argument (s_A, s) with $w = w_1(s_A, s) = w_2(s_A, s)$ and only the ratings and weights below (and including) s_A may differ in any way. No matter how those values are chosen, we want to have $d(T_1, T_2) \leq w$, i.e. the maximum influence of the differences below s_A is limited by the weight of s_A to its conclusion.*

Proof. It is enough to consider paths which include a, since the summands for all other parts are 0. Let S_a be the set of all statements which have an argument leading to a, including a itself. We abbreviate $\alpha^{\mathtt{depth}(s')}(1-\alpha)$ as $\overline{\alpha}(s')$.

$$d_T(T_1, T_2) = \sum_{s' \in S_a} \overline{\alpha}(s') \left| r_1(s') \prod_{s_{A'} \in \rho_{S \to I}} w_1(s_{A'}) - r_2(s') \prod_{s_{A'} \in \rho_{S \to I}} w_2(s_{A'}) \right| \tag{25}$$

$$= w \sum_{s' \in S_a} \overline{\alpha}(s') \left| r_1(s') \prod_{s_{A'} \in \rho_{S \to I} \setminus (s_A, s)} w_1(s_{A'}) - r_2(s') \prod_{s_{A'} \in \rho_{S \to I} \setminus (s_A, s)} w_2(s_{A'}) \right| \tag{26}$$

As the factor after w is in $[0, 1]$, we get $d_T(T_1, T_2) \leq w$. □

4. Limitations

Although the proposed pseudometric fulfills several intuitive desiderata, there are some limitations which we will discuss in the following.

If a weight of an argument is 0, the proposed pseudometric ignores all weights which are underneath this argument. When comparing how similar people argue for or against

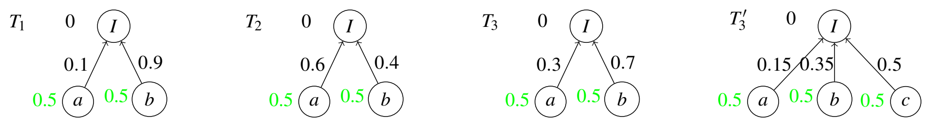


Figure 2. Possible unexpected change in order if an unrelated opinion is added: $0.05 = d(T_1,T_3) < d(T_2,T_3) = 0.075$, but $0.1375 = d(T_1,T_3') > d(T_2,T_3') = 0.125$ with $\alpha = 0.5$

the top-level positions, this is okay, but if also the way how arguments which are not supported are attacked should influence the distance, the metric has to be extended.

Moreover, ordering can be changed by adding an unrelated opinion. Consider trees T_1, T_2, T_3 with $d(T_1,T_3) < d(T_2,T_3)$. At first sight, it might seem unexpected that this order can be changed to $d(T_1,T_3') > d(T_2,T_3')$ by adding a new position c to T_3 and keeping the relative weights of the other positions. Depending on the application context, e.g. a voting advice application (VAA), this might be unwanted. An example is depicted in Figure 2. This is due to the normalization of the argument weights.

Although even end-user friendly systems like D-BAS support for undercuts, i.e. arguments that have an argument as conclusion, undercuts are currently not explicitly modeled in our model. This is no big problem, because in many applications, for instance, in a VAA, arguments can be preselected such that no undercut attack is necessary (since the arguments make sense), or rephrased such that the premise is attacked. For example, consider the argument “We should build more nuclear power plants because cats are cute” and the attack “Cats and nuclear power plants are unrelated”. Though this is technically an undercut, a user interface may present this as an attack on “Cats are cute”.

5. Related Work

Calculating the distance between argumentation graphs to compare how similar the attitudes of two agents are has already been used in other systems. The Carneades opinion formation and polling tool presented in [9] is able to compare one’s argumentation with the argumentation of other entities like organizations. This comparison is simply done by counting the number of statements where the agreement/disagreement is the same. This approach is much simpler than our proposal, but uses neither weights nor ratings and violates i.a., Desideratum 3.

The mobile application described in [1] also bases on IBIS and extends it with an agreement value for each argument in the argumentation tree. This information is used, for opinion prediction using collaborative filtering. In contrast to our work, the idea of relative argument importance in combination with statement rating is not present.

Another application of calculating the similarity of weighted trees is match-making of agents, which are represented by weighted trees. In [3], a recursive similarity measure for this application is proposed. Its parameter N serves a similar purpose to our parameter α. They also give examples which are similar to our desiderata, e.g. Example 4 is like our Desideratum 2. Nodes, however, do not have a weight, and some desiderata are not fulfilled; for instance, Desideratum 5 is explicitly not demanded in their Example 2.

There are already other definitions of weighted argumentation graphs based upon Dung’s definition of argumentation frameworks, but many lack the differentiation between argument relevance and statement ratings (e.g. [2,8,10,6]), which we think is important since argument weights limit how much a branch of an argumentation is relevant, whereas statement ratings are only relevant for the single statement. Furthermore, in most cases, there is a global assignment of values in the graph and no user-specific views, which is why a strength of 0 for attacks would be meaningless in the model presented in [13]. Note that most related work in this field is concerned about evaluating consistency or calculating extensions, whereas our main goal is comparing the attitudes of agents, not caring about whether an agents' attitude is logically consistent or not.

6. Conclusion and Future Work

In this paper we proposed a pseudometric to calculate the distance between two argumentation graphs representing the attitudes of different persons, and several desiderata which should be considered when proposing other metrics for the same purpose, and are fulfilled by our pseudometric. Possible next steps include developing other sensible metrics and comparing them regarding theoretical properties and practicality.

In future work, we want to check if the desiderata are not only intuitive for experts in argumentation theory, but are following the intuition of untrained humans. We also want to test the metric in a VAA to compare the argumentation of voters with those of parties. Thereby we see whether the results of the metric are accepted in an application context.

References

[1] Althuniyan, N., Sirrianni, J.W., Rahman, M.M., Liu, X.F.: Design of mobile service of intelligent large-scale cyber argumentation for analysis and prediction of collective opinions. In: International Conference on AI and Mobile Services. pp. 135–149. Springer (2019)
[2] Amgoud, L., Ben-Naim, J., Doder, D., Vesic, S.: Acceptability semantics for weighted argumentation frameworks. In: IJCAI. vol. 2017, pp. 56–62 (2017)
[3] Bhavsar, V.C., Boley, H., Yang, L.: A weighted-tree similarity algorithm for multi-agent systems in e-business environments. Computational Intelligence **20**(4), 584–602 (2004)
[4] Brenneis, M., Mauve, M.: deliberate – online argumentation with collaborative filtering. In: COMMA (2020), to appear
[5] Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (1995)
[6] Dunne, P.E., Hunter, A., McBurney, P., Parsons, S., Wooldridge, M.: Weighted argument systems: Basic definitions, algorithms, and complexity results. Artificial Intelligence **175**(2), 457–486 (2011)
[7] Ebbinghaus, B.: Decision making with argumentation graphs. arXiv preprint arXiv:1908.03357 (2019)
[8] Gordon, T.F., Walton, D.: Formalizing balancing arguments. In: COMMA. pp. 327–338 (2016)
[9] Gordon, T.F.: Structured consultation with argument graphs. From Knowledge Representation to Argumentation in AI. A Festschrift in Honour of Trevor Bench-Capon on the Occasion of his 60th Birthday pp. 115–133 (2013)
[10] Hunter, A.: A probabilistic approach to modelling uncertain logical arguments. International Journal of Approximate Reasoning **54**(1), 47–81 (2013)
[11] Krauthoff, T., Meter, C., Mauve, M.: Dialog-Based Online Argumentation: Findings from a Field Experiment. In: Proceedings of the 1st Workshop on Advances in Argumentation in Artificial Intelligence. pp. 85–99 (November 2017)
[12] Kunz, W., Rittel, H.W.J.: Issues as elements of information systems, vol. 131. Citeseer (1970)
[13] Martınez, D.C., Garcıa, A.J., Simari, G.R.: An abstract argumentation framework with varied-strength attacks. In: Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR'08). pp. 135–144 (2008)
[14] Neugebauer, D.: Dabasco: Generating af, adf, and aspic+ instances from real-world discussions. In: COMMA. pp. 469–470 (2018)

3.2.3 Further Thoughts on a Metric

When developing the distance measure, an important design decision was that it should fulfill basic properties of a metric, especially the triangle inequality. These properties allow taking advantage of existing optimizations for neighborhood-based recommender systems. The optimizations use the triangle inequality to prune unnecessary calculations, thus accelerating the recommender system. An example are M-trees (Ciaccia et al., 1997) for efficient k-nearest neighbor queries.

There is another aspect which should be considered when applying a distance function like ours which works on the structure of the graph, but does not consider the actual content of an argument. In case arguments which are semantically very similar are present, the difference between a pair of graphs might be overestimated. In the context of assessing the strength of attacking arguments, this problem was investigated by Amgoud and David (2020), who introduced a so-called adjustment function to account for similarity of arguments.

Another limitation already touched on in the paper is the possibility of "manipulation" by changing the resulting order by adding unrelated opinions: This problem is actually quite similar to the condition of independence of irrelevant alternatives by Arrow (1950) in the context of social choice theory. In brief, Arrow's impossibility theorems states that no ranked voting electoral system exists which assures all of these properties:

- If everyone prefers X over Y, the group prefers X over Y.
- If the group prefers X over Y, any preference change not involving the order of X and Y does not change the group's preference.
- The group's preference cannot be determined by a single person.

In fact, similar assertions could be applied to the comparison of position weights, which are a kind of ranking, and where the voting outcome is the question of which weighted argumentation graph is closest to most other graphs. We will get back to the question of order manipulation when talking about VAAs in Subsection 4.2.2.

3.3 A Human Baseline for Comparing Argumentations

In Section 3.2, we developed a distance function which enables us to calculate how similar the attitudes of two persons are in an argumentative context. We assembled a list of desiderata which are fulfilled by our pseudometric to ascertain that it follows intuitive properties. But how can we be sure that those properties are actually intuitive not only for us, but also for average humans, who do not deal with argumentation theory every day? Would an average person actually apply those "rules" themself when doing a comparison of argumentations? If a metric is supposed to yield results which should be understood, in particular in sensitive application like a VAA, it is essential to have a justification that its results are plausible for most people.

Therefore, we conducted a survey to find out how human subjects unfamiliar with argumentation theory assess the similarity of argumentations. In this section, we present the design of and key findings from our empirical study. We start off with an overview of our hypotheses we wanted to verify with the survey. Subsequently, we delve into our main results, which we already previously published (Brenneis and Mauve, 2020b) and contain some surprising discoveries; for instance, “neutral” was not considered to fall on a line between “pro” and “contra” by many people. We then elaborate on the hypothesis test used for evaluating the study, and finally look at the relation of our desiderata and our survey results.

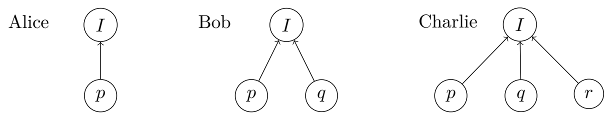
3.3.1 Overview of Our Hypotheses

We were interested to know whether the desiderata from Subsection 3.2.2 are matching human intuitions, and if the limitations we identified are actual limitations. So we developed different argumentation scenarios based on our desiderata, the possible limitations of our pseudometric, and further questions about, for example, trade-off situations. We came up with 22 main hypotheses in four categories, which we will expound on now. Every scenario is accompanied by a visualization to clarify the attitudes involved.

Basics

We first focus on basic properties of argumentation graphs, i.e. having opinions for/against a different number of statements and adding arguments to positions and other statements.

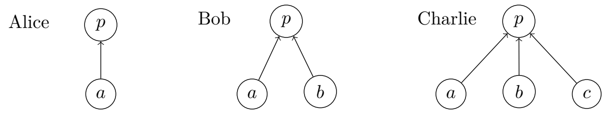
When persons give their opinions on a set of positions, are their attitudes considered to be more similar if the relative overlap of their opinions is greater or is the absolute number of differences more important?



For example, Bob has the same opinion on a position p as Alice, but gives an additional opinion on another position q. Charlie has the same opinions as Bob, but gives an additional opinion on position r. We think that Bob and Charlie are considered closer to each other since their proportional overlap of opinions is greater; we already had this anticipation for Desideratum 1 in Subsection 3.2.2. On the other hand, one could assume that only the absolute number of differences is important, i.e. Bob is equally far apart from Charlie and Alice because Bob has one more position than Alice and Charlie one more position than Bob.

Hypothesis 1 *Proportionally bigger overlap of opinions on positions results in greater similarity than the absolute number of differences.*

What we asked for positions in Hypothesis 1 can also be asked for arguments:



Consider that Bob has the same opinion on a position p as Alice, but gives more arguments, and Charlie has also the same opinions, but mentions more arguments than Bob. We think that Bob and Charlie are considered closer since their proportional overlap of arguments used is greater: Bob and Charlie overlap in 2 of 3 arguments, but Alice and Bob only in 1 of 2 arguments. On the other hand, one could assume that only the absolute number of differences is important, which is 1 when comparing Alice with Bob, and Bob with Charlie.

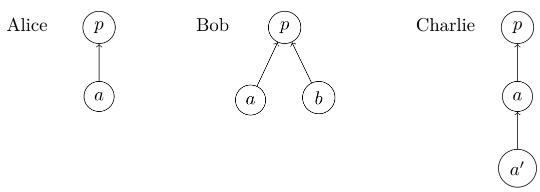
Hypothesis 2 *Proportionally bigger overlap on arguments for/against a position results in greater similarity than the absolute number of differences.*

How does a neutral opinion influence the distance?

If Alice, Bob, and Charlie are for, against, and neutral to a position, respectively, we expect that Charlie is right between Alice and Bob, i.e. Charlie's distance is the same to Alice and Bob, and Alice and Bob are further away from each other than Alice and Charlie. This assumption is based on Desideratum 2.

Hypothesis 3 *A neutral opinion is between a positive and a negative opinion.*

Is the level where arguments are added relevant?



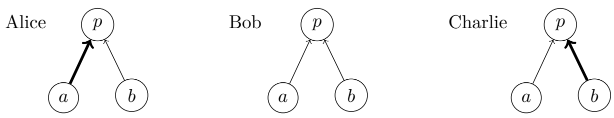
This scenario is also described in our paper (Brenneis and Mauve, 2020b), at which we look in Subsection 3.3.2. Alice, Bob, and Charlie have the same opinions concerning a position p and a common argument a for p. Now Bob adds another argument for p and Charlie an argument with conclusion a. We anticipate that Alice and Charlie are closer to each other than Alice and Bob because their first-level argumentation is the same and the difference is in a deeper part; we already conjectured this in Desideratum 3. However, one could also conjecture that individuals who are not familiar with argumentation theory do not have a feeling for the different levels and therefore regard both differences in the argumentation behavior as similarly big.

Hypothesis 4 *Deviations in deeper parts have less contribution to dissimilarity than deviations in higher parts.*

Influence of weights/importance

The following hypotheses focus on the question of how weights for opinions and arguments play a role.

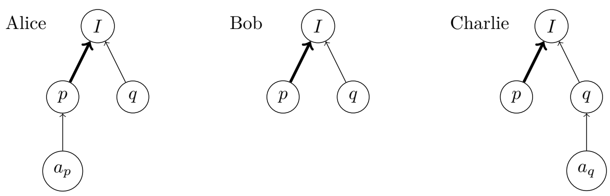
If two persons have the same opinions and arguments regarding a position, but they give different weights to the arguments, are those weights considered or are only the identical opinions considered important?



Is there no perceived differences in attitude if the only difference are the weights? We expect the different weights to have an influence on the perceived difference, as already described in Desideratum 5.

Hypothesis 5 *Weights of arguments have an influence even if they are the only difference.*

Do deviations within argumentation branches of less important arguments have a smaller influence?



Consider Alice, Bob, and Charlie have the same opinions on two positions p and q and consider p more important than q. Now Alice gives an argument for p and Charlie an argument for q; Bob does not mention these arguments. We expect that Bob is considered closer to Charlie than to Alice, as their argumentation differences are in a less important branch of the argumentation. On the other hand, one could also expect that humans do not consider the different importances of the branches and only count the number of arguments given.

Hypothesis 6 *Argumentation differences in a branch with lower importance contribute less to dissimilarity.*

This hypothesis matches our Desideratum 4.

Influence of missing information/no opinion

The following hypotheses are related to missing information: In a real-world application, not all opinions and weights of an individual person are known. How should those cases be treated?

Is indicating to have no opinion the same as being neutral?

Alice $+\ p$ Bob $-\ p$ Charlie $\emptyset\ p$

If Alice and Bob are for or, respectively, against a position, and Charlie indicates to have no opinion, we expect that Charlie is right between Alice and Bob. This means that Charlie's distance is the same to Alice and Bob, and Alice and Bob are further away from each other than Alice and Charlie (similar to Hypothesis 3, but Charlie has no opinion instead of being explicitly neutral).

Hypothesis 7 *No opinion is between a positive and a negative opinion.*

One could argue, though, that information is missing to give a good assessment. A complete mathematical metric, however, has to make some kind of decision. In some application contexts, it might make sense to have an incomplete metric yielding an undefined distance, but this is certainly not true for every application. Therefore, when checking this hypothesis, we want to test two cases: In the first case, a decision has to be made by participants. In the second case, not making a decision is allowed.

Can an unknown opinion be treated the same way a neutral opinion is treated?

Alice $+\ p$ Bob $-\ p$ Charlie $?\ p$

If Alice is for and Bob against a position, and Charlie's opinion is not known, we expect that Charlie is right between Alice and Bob, i.e. Charlie's distance is the same to Alice and Bob, and Alice and Bob are further away from each other than Alice and Charlie (similar to Hypothesis 3, but Charlie's opinion is not known instead of being explicitly neutral).

Hypothesis 8 *An unknown opinion is between a positive and a negative opinion.*

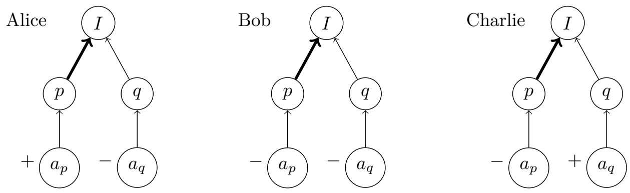
As in Hypothesis 7, forcing a decision and allowing making no decision has to be tested.

How is an opinion which is not mentioned treated? Not mentioning an opinion should have the same effect as explicitly saying that an opinion is unknown.

Hypothesis 9 *A statement for which no opinion is mentioned is like a statement for which we explicitly say the opinion is unknown.*

This hypothesis is not really supposed to bring new insights for the development of a metric, but to assure that the text comprehension is in line with our thoughts. In other hypotheses, we often have the situation that certain opinions are not mentioned, and we want to ascertain that this is treated like explicitly saying that the opinion is unknown.

Do not mentioning an argument and being against an argument have the same effect?



Consider Alice, Bob, and Charlie have the same opinions on two positions p and q and think that p is more important than q. Now Alice gives an argument with premise a_p for p, and Charlie an argument with premise a_q for q; Alice does not accept the premise a_q, Charlie does not accept the premise a_p, and Bob does not accept both premises (this setting is similar to Hypothesis 6, but instead of not mentioning an argument, the persons are against the premise).

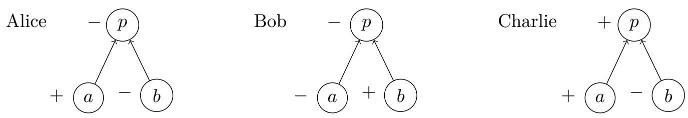
We expect that not mentioning an argument (as in Hypothesis 6) is the same as being explicitly against that argument (as here) if the opinion on the conclusion is the same. So in this scenario, we also expect that Bob is considered closer to Charlie as their argumentation differences are in a less important branch.

Hypothesis 10 *Not mentioning an argument and being against an argument have the same effect.*

Trade-offs

Different properties of the argumentation graphs probably have different influences on the perceived difference. But how strong, for example, is the influence of the opinion compared to the arguments used? Therefore, we also contemplate trade-off situations.

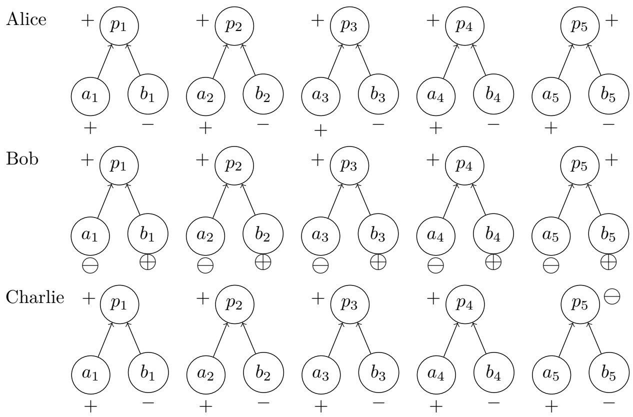
Are the opinions on a position or the arguments agreed to more important?



Consider that Alice and Bob have the same opinion on a position, but the arguments they use are contradictory. Charlie has the same opinions regarding the arguments as Alice, but her opinion on the position is different. We expect that Alice is closer to Bob since people probably consider the opinions on the positions more important. On the other hand, it would be reasonable to say that Alice and Charlie are closer, because they agree on all arguments, and only differ in their conclusion.

Hypothesis 11 *Disagreeing on a position results in greater distance than having the same opinion on that position, but with contrary arguments.*

Is it possible that disagreement on arguments outweighs differences in opinions?



Basing upon Hypothesis 11, we want to know whether it is possible to construct an extreme situation where a differing opinion is outweighed by the use of arguments. We think that not only the agreement on an opinion for a position is considered when assessing the similarity of argumentation behavior, but also the arguments used.

So if Alice and Bob agree on many positions, but disagree on their arguments, Charlie is considered more similar to Alice if Charlie agrees with all arguments of Alice, but has few differing opinions (for better readability, differences from Alice's attitudes are encircled in Bob's and Charlie's graphs). It would also be reasonable, though, to assume that similarity is determined step-by-step, i.e. only the top-level opinions are looked at, and arguments are only considered if those opinions differ.

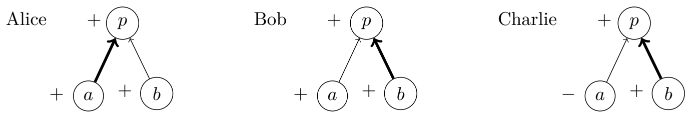
Hypothesis 12 *It is possible for a difference in arguments for/against positions to result in greater dissimilarity than a difference in opinions on those positions.*

Is it possible that an argumentation is more similar to an argumentation with opposite opinion, if the opinions in both argumentations are weak enough?

We expect that a very big difference in the weights for different positions can lead to the effect that persons with contrary, but weak opinions are considered more similar than persons who have the same opinion, but one opinion is strong and the other very weak. In this example, this means that Bob would be seen closer to Charlie than to Alice. This expectation is based on Desideratum 8. It would also make sense to assume that only the opinion tendency is relevant and the strength of the opinion is not considered.

Hypothesis 13 *Two argumentations with weak and contrary opinions on a statement can be closer than two argumentations with the same opinions, but with very different strengths.*

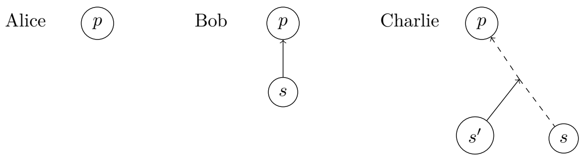
We expect a similar trade-off effect for arguments:



It is more important to agree on the position and the main argument than having an opposing view on an unimportant argument. Hence, we anticipate that Bob and Charlie are closer to each other than Bob and Alice in the example depicted above, as already expounded on in Desideratum 7.

Hypothesis 14 *Two argumentations with weak arguments and contrary opinions on their premises can be closer than two argumentations with the same opinions, but with very different strengths of arguments.*

Do opinions which are a premise of an undercut argument count towards a person's attitude?



Let us assume Alice, Bob, and Charlie have the same opinion regarding a position p. Bob gives an additional argument $a = (s, p)$ with premise s for p, but Charlie claims that a is not related to p, although she accepts s (undercut attack). We expect that opinions below a not accepted argument are irrelevant when determining the differences in attitude regarding a position, i.e. there is no perceived difference between Charlie and Alice, although Alice's opinion on s is unknown. This expectation implies that the limitation of our pseudometric, which ignores branches with a weight of 0 as pointed out in Subsection 3.2.2, is no real limitation. One could, however, also assume that those additional arguments are considered part of the attitude to the position.

Hypothesis 15 *When determining the attitude regarding a position, opinions (not) mentioned for a not-accepted argument have no influence.*

Persons can have different relevances for their positions. How should a metric treat those values?

Alice:	Bob:	Charlie:
1. b	1. a	1. a
2. a	2. c	2. b
3. c	3. b	3. c

Do individuals compare only the top positions of a ranking? Consider Alice, Bob, and Charlie have the same opinions on several positions, but consider them of different importance. If Alice has the same priorities as Charlie, but swaps the most important positions in order, and Bob the same priorities as Charlie, but swaps two less important positions, we expect Charlie to be considered closer to Bob than to Alice. One could expect a different outcome if similarity is assessed step-by-step, meaning only the two most important positions of each person are looked at.

Hypothesis 16 *Flipping the two important positions results in a bigger difference than flipping two less important positions.*

What happens if someone adds a position in their order which is not mentioned by anyone else?

Alice:	Bob:	Charlie:	Charlie':
1. b	1. a	1. a	1. d
2. a	2. c	2. b	2. a
3. c	3. b	3. c	3. b
			4. c

Is it possible to remove a previous dissimilarity this way? Basing upon Hypothesis 16, we wanted to know what happens if Charlie adds a new most important position which is not mentioned by the other persons. Our expectation is that Charlie' is considered equally far away from Alice and Bob because all other positions are now of less importance. Another reasonable expectation would be that this new position is ignored, i.e. the results are the same as in the scenario for Hypothesis 16.

Hypothesis 17 *Adding a new position can remove a previous dissimilarity.*

Going even further than in the previous case, is it possible to swap a similarity order by adding a new position?

Alice:	Bob:	Charlie:	Charlie':
1. a	1. d	1. a	1. e
2. c	2. a	2. b	2. a
3. d	3. b	3. c	3. b
4. b	4. c	4. d	4. c
			5. d

Similar to Hypothesis 17, we think that it is possible to construct a scenario where adding a new, fifth position which is considered most important can swap distances. For the example given above, one can think that Charlie is closer to Alice before adding the fifth position and closer to Bob after adding it. This conclusion can be drawn by counting the number of absolute place differences for each common statement (Charlie–Alice: 4, Charlie–Bob: 6; Charlie'–Alice: 6, Charlie'–Bob: 4).

Hypothesis 18 *Adding a new position as most important position can swap a previous similarity order.*

We introduced this hypothesis to check whether the limitation of our metric which causes a change of order when adding a new position (discussed in Subsection 3.2.2) is a real problem.

How do priorities and opinions play with each other?

Alice:	Bob:	Charlie:
1. $a+$	1. $b+$	1. $b-$
2. $b+$	2. $a-$	2. $a+$

Is it more important that my most important opinions match yours or that your most important opinions match my opinion? Consider the following setting: Alice's opinion on her top position is not in agreement with Bob, but Bob also puts this position on the last place. The contrary is true for Charlie: She disagrees with Alice's last position, but Charlie has it as most important position.

We think that Bob and Charlie are equally far apart from Alice's attitude, as the number of dissimilarities is equal: If you want to change Alice's list to Bob's, you change the top opinion and do a swap; you do the same when changing Charlie's list to Alice's. Another reasonable assumption would be that Alice is considered to be more similar to Charlie, since Charlie matches Alice's most important positions. The latter is a person-centric, asymmetric interpretation.

Hypothesis 19 *Agreeing with someone's most important position is as important as having that person's most important opinion matching mine.*

How does adding a new top-1 position influence the difference in comparison to flipping the priorities of two positions?

Alice:	Bob:	Charlie:
1. a	1. b	1. c
2. b	2. a	2. a
		3. b

For instance, Alice agrees with the positions a and b, where a is more important for her. Bob also agrees, but swaps the priorities. Charlie has the same order as Alice, but another, most important position c. We think that flipping the order of two (same) opinions is not as severe as adding a new most important opinion, since a and b have lower priorities for Charlie than for Bob. One could also argue, though, that Charlie and Alice are closer because their order is the same.

Hypothesis 20 *Adding another most important position results in greater dissimilarity than flipping the priorities of two positions.*

Can priorities for positions be more important than the number of same opinions?

Alice:	Bob:	Charlie:
1. a	1. a	1. c
2. b		2. b
3. c		

Consider Alice, who agrees with the positions a, b, and c with priorities in this order. Bob only has a positive opinion for a, which is most important for him. Charlie considers c and b most important (in this order). Although Alice and Charlie have a greater overlap in opinions, we think the top-priority of c makes Alice more similar to Bob than to Charlie. This means we do not think that individuals only count the number of agreements, but also consider the relevance of the positions.

Hypothesis 21 *Having more similar priorities of opinions can result in greater similarity even with lower absolute number of same opinions.*

Does considering all positions weigh more than decreasing priorities by adding another point?

Alice:	Bob:	Charlie:
1. a	1. b	1. c
2. b		2. b
3. c		

Consider Alice, who agrees with the positions a, b, and c with priorities in this order. Bob only has a positive opinion for b, which is most important for him. Charlie considers c and b

most important (in this order). We think that Charlie's attitude is closer to Alice's attitude, because Alice considers all position of Charlie, albeit with lower and swapped priorities. It would also make sense to think that Charlie is closer to Bob because Alice has higher priorities for positions which Charlie does not mention (i.e. position a).

Hypothesis 22 *Not mentioning a position results in greater dissimilarity than assigning lower priorities.*

Desiderata Which Are not Covered

Two desiderata have no (directly) equivalent hypotheses.

Desideratum 6 (*Symmetry regarding negation of opinion*) was not included. First of all, this fundamental property should be followed in any sensible comparison. Secondly, we do not really have to check this explicitly, as we already do this implicitly with the questions we ask about other scenarios. We asked for every possible similarity assessment combination in most cases, thereby getting information on the symmetry. Lastly, symmetry is required for any metric, and we want to obey the basic laws of metrics as discussed in Subsection 3.2.3.

Desideratum 9 (*Weights limit the influence of a path*) remained untested as this property was considered too mathematically and untestable without leaking the model of numeric weights into the questionnaire scenario. The questionnaire should be agnostic with regard to the underlying mathematical model.

3.3.2 Details on Our Methods and Key Findings

Now we have a look at which hypotheses turned out to match human intuition, which did not, where no clear results could be found, and which methods we used to get those findings. The key results of our survey have been published in the following peer-reviewed conference paper:

Markus Brenneis and Martin Mauve.
"Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations"
In: *Proceedings of the 4th Workshop on Advances in Argumentation in Artificial Intelligence (AIˆ3 2020)*, *AI*IA Series*, pages 1–15, CEUR Workshop Proceedings Volume 2777.
Acceptance Rate: 89%

Within this publication, which is included in this section, we presented the following information and results:

1. our hypotheses for the assessment of attitude similarity in argumentations
2. the transformation of hypotheses to questionnaire scenarios

3. the design of our empirical study
4. results for the assessment of the scenarios by untrained human subjects
5. an interpretation of the results

We conducted our survey using *Amazon Mechanical Turk* (MTurk) with citizens from the US because of the fast and cheap recruiting process. We originally wanted to survey German-speaking people through MTurk, since we already had in mind to apply the results in the construction of a German argument-based VAA, which will be presented in Section 4.2. This try failed, though, since only very few German people participated in the survey, even for small questionnaires and high incentives. Using a German survey institute would have been more expensive. On-campus recruitment was not possible due to the Covid-19 pandemic at that time and would have been quite unrepresentative.

In the end, around 40 answers per hypothesis could be collected. Out of 33 hypotheses (including sub-hypotheses), we could significantly confirm 19. Plausible, alternative answers were found for 5 hypotheses.

Personal Contribution

Markus Brenneis, the author of this thesis, developed the idea for surveying untrained people regarding their assessment of argumentation similarity and the list of hypotheses. He designed drafts for the scenarios in the questionnaire, conducted the survey, analyzed the results, and wrote the whole paper. Martin Mauve dispensed advice regarding the structure of the paper. The questionnaire design and the results were discussed with Martin Mauve, who gave useful input for clarifying some questions and hypotheses. We want to thank the students Jan Steimann, Marc Feger, Lian Remme, Miriam Detlefs, and Tim Neumann for their feedback on early versions of the questionnaires and linguistic mistakes. The design of a suitable hypothesis test was developed by Markus Brenneis and was refined through several conversations with Holger Schwender, who gave useful keywords for finding relevant literature.

Importance and Impact on This Thesis

This paper continues the work on our pseudometric presented in Section 3.2, since it lies the foundation for justifying why the desiderata, and hence, our pseudometric, are close to human intuition. Moreover, the results of the survey were expected to give hints on whether the limitations of the pseudometric discussed in Subsection 3.2.2 are actual issues. This thought is completed by the work we will present in Section 3.4, where we use the results of this survey to compare different possible distance functions and their parametrizations regarding their intuitiveness. Having an intuitive distance function is important for justifying its use in applications like our VAA (cf. Section 4.2) where the results should be understandable to everyone.

Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations*

Markus Brenneis and Martin Mauve

Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany
Markus.Brenneis@uni-duesseldorf.de

Abstract. In this paper, we present the results of a study where participants were asked to rate the similarity between sets of positions and arguments. Our goal is to provide a baseline for metrics that compare the attitudes of individual persons in argumentations, with results matching human intuition. Such metrics have different applications, i.a. in recommender systems. We formulated several hypotheses for useful properties, which we then investigated in our survey. As a result, we were able to identify several properties a metric for comparing attitudes in argumentations should have, and got some surprising results we discuss in this paper (e.g., many people do not see a "neutral" position on a line between "pro" and "contra"). For some properties, further research is needed to get a clearer understanding of human intuition.

Keywords: Argumentation · Metric · Human Baseline.

1 Introduction

When discussing with other people, it is interesting to know how similarly another person argues like yourself, i.e. how similar your attitudes are. Do you disagree on central statements, or do you generally agree, but differ in some arguments? Do you have the same priorities for political positions or the same reasons, e.g. for the expansion of wind power? Having a mathematical metric for calculating the (dis-)similarity of attitudes in argumentation enables use-cases like collaborative filtering for argumentation applications like *kialo*[1] or our *deliberate* [5], finding representatives of a group, finding a consensus, and matching political parties and voters based on attitudes and used arguments.

People typically discuss central positions (e.g. the improvement of a course of study [12] or the distribution of funds [8]) and support (or attack) them with other statements, which we call an argument. Each individual person agrees or disagrees more or less strongly with certain statements, and may consider some arguments more important than others when forming an opinion.

When designing a metric for an application where arguments are exchanged, one has to ask which properties that metric should fulfill. For instance, should an

[1] https://kialo.com/

opinion difference in "top-level" arguments against a position (e.g "We should not build nuclear plants, because they are insecure") weigh more than disagreement on "deeper" arguments (e.g. "Nuclear plants are insecure, because there have been several accidents.")? Are two persons who are against and for a position equally far apart from each other as two persons where one is for a position, and the other one has a neutral opinion? (Surprisingly for us, our results indicate that the latter is, in fact, the case, as we will explain in Section 4.2.)

Any reasonable metric to answer those questions needs to be based on the perception that humans have regarding the similarity of chains of arguments, instead of the "intuition" of researchers who deal with argumentation theory every day. To establish a baseline for this, we asked our survey participants to judge the similarity of two chains of argumentation. Which pair is considered more similar? The questions asked were based on hypotheses presented in this paper. The hypotheses should help with answering how a metric should behave in trade-off situation, with missing information, hierarchies, and weights in argumentations. To our knowledge, such a survey has not been conducted before.

Our contribution is the following: We formulate several hypotheses for assessing the similarity of argumentations, which should be respected by a metric comparing attitudes expressed in argumentations. We gathered a data set with human assessments of relative similarity of argumentations for testing the real-world relevance of our hypotheses, and checked which hypotheses can be regarded as correct with a high significance.

In the following section, we define central concepts of argumentation theory relevant for this paper. Afterwards, we describe our methods used and our hypotheses. We then present our most important and surprising results. In the fifth section, we discuss our methods, and finally, we comment on related work.

2 Definitions

In this paper, we use terms based on the IBIS model [13] for argumentation. Within an argumentation context, there are *arguments*, which consist of two *statements*: a *premise* and a *conclusion* (e.g., "Nuclear power is sustainable." can be a premise for "We should build a nuclear power plant."). When we draw an argumentation graph, statements are nodes, arguments are edges. Statements which are only used as conclusion are called *positions*, and are typically actionable items like "We should build a nuclear power plant". The unique root of the argumentation graph is called *issue* I, and connects all positions. It is typically the overall topic of the discussion, e.g. "What shall the town spend money for?".

Each person can have a specific view on the parts of an argumentation graph: A person can agree or disagree with a statement, which we call the person's *opinion*. Arguments and statements can be of different importance (or relevance, weight) to different persons. Each individual person may use one specific subset of all available arguments. We call the sum of opinions, importances, and arguments used by a person *attitude*.

The results of our work are independent of this model, but it enables us to precisely formulate our hypotheses (i.a. by having statements, not arguments, as atomic elements), and draw graphs for visualizing scenarios for our hypothesis. So our findings can also be applied to metrics working with Dung-style [7] argumentation frameworks; for instance, our issue-based graphs can be transformed to an abstract argumentation framework using the tool *dabasco* [16].

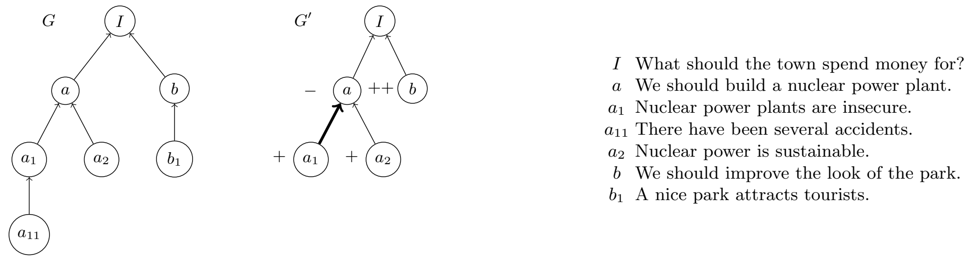


Fig. 1: Example for an argumentation graph G and a personal view G' on that graph G with attitudes. Statements with unknown opinion are not drawn in G'.

To understand how our graphs should be read, Figure 1 depicts an example of an argumentation graph G for a discussion and a personal view G' on that graph, which contains Alice's attitudes. In this example, Alice is very sure (++) that she wants the look of the park being improved (b), and she is against a nuclear power plant (a, $-$). She accepts the statements that nuclear power is sustainable (a_2) and nuclear power plants are insecure (a_1), but she thinks the latter weighs more (thick line) for her opinion on building a nuclear power plant. Alice has not mentioned an opinion on the statements a_{11} and b_1.

We will not draw opinions for better readability if the focus of a scenario is not on opinions, and they are considered to be the same across graphs being compared (e.g., “agree”/“+” can be assumed for all statements in Figure 2).

3 Methods

We now present how we developed our hypotheses for properties of a metric for comparing the way different persons or organizations argue, how we created questionnaire scenarios, and conducted the survey. Our focus is explicitly on comparing the *attitudes* of different persons within an argumentation, not properties like number of counterarguments, consistency, or use of rhetorical devices.

We are well aware that our list of properties is only a starting point for the work of finding out how human feeling of argumentation similarity can be translated to a mathematical metric. Thus, we expect that our list can be extended with more properties in the future.

First, we formulate hypotheses about what we expect of a metric. Those hypotheses are at least somewhat reasonable for domain experts, and are partially based on properties of a metric we have presented in an earlier work [4]. However, before they are used for guiding the development of metrics for the comparison of argumentations, it should be checked whether they match the perception of average humans.

To do so, we developed questionnaire scenarios for every single hypothesis. Participants of the survey were asked to assess the similarity of the people's argumentation by indicating which person's argumentation is most similar to the argumentation of another given person. For scenarios which involved only one topic (e.g. an argumentation on nuclear power), we had multiple versions of that scenario with different topics to prevent topic-dependent results.

The survey was conducted using Amazon Mechanical Turk (MTurk) because of its easy and fast recruiting process. Only participants from the US were allowed to assure that there is a sufficient knowledge of English. Although MTurk users are not representative for the US population, it has been shown that the average difference can be quite small [2]. The questions and scenarios were randomly assigned to the participants and the order of answers was randomized. To assure answers of good quality, only answers of participants who answered at least 3 of 5 quality control questions correctly were used in the evaluation.

The complete list of hypotheses is in Table 1. They are grouped in four categories with different motivations: First, we were interested in the influence of basic properties of argumentations, like being for/against a different number of statements and adding arguments. Then we asked ourselves what the influence of weights of opinions and arguments is, and whether they play a role at all. The third group deals with the influence of missing information: Real-world applications often do not have complete information of a person's attitude, how should a metric behave here? The last is about trade-off situations: What weighs more when both, opinions and arguments mentioned, are different between persons? What is the influence if the relevance of positions is rated completely different?

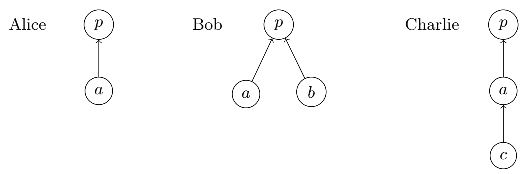


Fig. 2: Visualization of the scenario for Hypothesis 4: The graphs represent the attitudes in the argumentation of each person in the scenario.

As an example, we now present how Hypothesis 4 (*deviations in deeper parts have less contribution to dissimilarity than deviations in higher parts*) has been developed and transformed in a questionnaire scenario. All scenarios can be found in our complete data set which is available online.[2]

[2] https://github.com/hhucn/argumentation-similarity-survey-results

We asked ourselves whether the level where arguments are added is relevant. To make the idea of the hypothesis clearer, Figure 2 depicts the attitudes of the persons involved in the constructed scenario.

Consider Alice, Bob, and Charlie have the same opinions on a position p and a common argument a for it. If Bob adds another argument for p, and Charlie an argument to a, we think that Alice and Charlie are closer because their first-level-argumentation is the same and the deviation is in a deeper part. One could, however, also assume that individuals not familiar with argumentation theory do not have a notion for levels and consider both differences in argumentation behavior as similarly severe.

From our hypotheses, we constructed the following scenario and questions:

> Alice argues as follows on the subject of wind power:
> More wind turbines should be built because wind power has a **low environmental impact**.
> Bob argues as follows:
> More wind turbines should be built because wind power has a **low environmental impact** and because wind turbines are **safe**.
> Charlie argues as follows:
> More wind turbines should be built because wind power has a **low environmental impact**. The **reason for the low environmental impact** is that they do **not produce any emissions**.
>
> Whose attitude does Alice agree with most?
> – with Bob's attitude – with Charlie's attitude – the attitudes are equally far apart
> Whose attitude does Bob agree with most?
> – with Alice's attitude – with Charlie's attitude – the attitudes are equally far apart
> Whose attitude does Charlie agree with most?
> – with Alice's attitude – with Bob's attitude – the attitudes are equally far apart

The relevant question for us is *Whose attitude does Alice agree with most?* and our expected answer is *with Charlie's attitude*; the other questions were added for gathering additional data and preventing biased answers.

Most other scenarios are constructed the same way. An exception are questions related to missing information, where we asked the questions twice: Once we forced a decision (since a complete, well-defined metric has to make some decision, too), and once we allowed to choose *this cannot be assessed* as an answer.

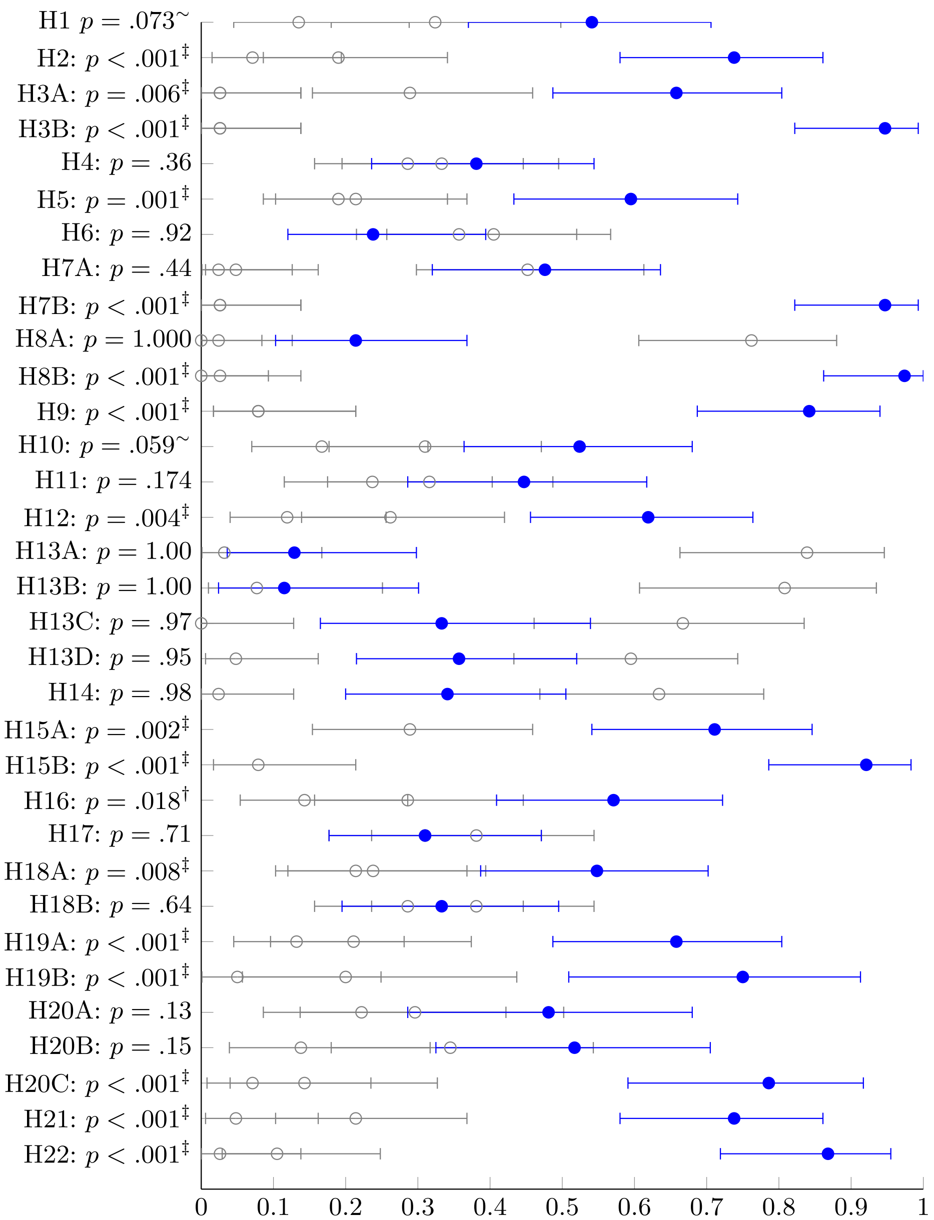
4 Results

We now present the results of our survey, and highlight and explain results which were surprising for us. We report p-values for the null hypothesis "our expected answer is not the most frequently (relative frequency) given answer".[3] For space reasons, not all numbers are presented and discussed in detail, but the aggregated raw data for all questions is available online. A summary of the relative answer frequencies for the relevant questions is depicted in Figure 3.

[3] We used an intersection–union test [18, p. 240] with one-tailed tests on the variances of the difference of two multinomial proportions [9,17], i.e. H_0 is that the differences of the relative answer frequencies between the expected answer and the other answers is not greater than 0.

Table 1: Our hypotheses about the assessment of attitude similarity in argumentations, grouped in basic properties, influence of weights, influence of missing information, and trade-offs

#	Hypothesis
H1	Proportionally bigger overlap of opinions on positions results in greater similarity than the absolute number of differences.
H2	Proportionally bigger overlap on arguments for/against a position results in greater similarity than the absolute number of differences.
H3	A neutral opinion is between a positive and a negative opinion.
H4	Deviations in deeper parts have less contribution to dissimilarity than deviations in higher parts.
H5	Weights of arguments have an influence even if they are the only difference.
H6	Argumentation differences in a branch with lower importance contribute less to dissimilarity.
H7	No opinion is between a positive and a negative opinion.
H8	An unknown opinion is between a positive and a negative opinion.
H9	A statement for which no opinion is mentioned is like a statement for which we explicitly say the opinion is unknown.
H10	Not mentioning an argument and being against an argument have the same effect.
H11	Disagreeing on a position results in greater distance than having the same opinion on that position, but with contrary arguments.
H12	It is possible for a difference in arguments for/against positions to result in greater dissimilarity than a difference in opinions on those positions.
H13	Two argumentations with weak and contrary opinions on a statement can be closer than two argumentations with the same opinions, but with very different strengths.
H14	Two argumentations with weak arguments and contrary opinions on their premises can be closer than two argumentations with the same opinions, but with very different strengths of arguments.
H15	When determining the attitude regarding a position, opinions (not) mentioned for a not-accepted argument have no influence.
H16	Flipping the two most important positions results in a bigger difference than flipping two less important positions.
H17	Adding a new position can remove a previous dissimilarity.
H18	Adding a new position as most important position can swap a previous similarity order.
H19	Agreeing with someone's most important position is as important as having that person's most important opinion matching mine.
H20	Adding another most important position results in greater dissimilarity than flipping the priorities of two positions.
H21	Having more similar priorities of opinions can result in greater similarity even with lower absolute number of same opinions.
H22	Not mentioning a position results in greater dissimilarity than assigning lower priorities.



Clopper–Pearson confidence intervals ($\alpha = 0.05$) indicated with expected answer (blue, filled circles) and other answer options (gray) for the relevant question, p-value for H_0 “expected answer is not the most frequently given answer”, ‡: $p \leq 0.01$, †: $p \leq 0.05$, $\sim$: $p \leq 0.10$

Fig. 3: Results for the relevant questions for each hypothesis

After removing participants who did not meet our quality standards, we had, on average, 38 answers for every question relevant for our hypotheses. Those participants have a median age of 30-39 years, which matches the US median of 2018 (36.9). The male/female ratio is 1.96 (total US ratio 0.97), thus we had significantly more male than female participants in our random MTurk sample.

4.1 Results that confirmed our expectations

For many scenarios, we did not get surprising results, and summarize them here.

Proportionally bigger overlap of arguments (H1) or opinions (H2) is indeed more important than the absolute number of differences (H1: expected answer given by 54%, $p = .073$; H2: 74%, $p < .001$). If the assessment of argument relevance is the only difference between attitudes, this is considered as difference by most participants (H5, 60%, $p = .001$).

That the most important opinion in one argumentation matches the opinion in the other argumentation is as important as the reverse case (H19), independent on whether this questions is asked from a person-centric (66%, $p < .001$) or "bird's eye view" (70%, $p < .001$). Flipping the priorities of the most important positions results in a smaller perceived difference than adding a new most important position p, regardless of whether the other persons have not mentioned their opinion on p (H20A, 48%, $p = .13$), had an explicit unknown opinion (H20B, 52%, $p = .15$), or were neutral (H20C, 79%, $p < .001$). Leaving out a position results in a greater dissimilarity than lowering its priority (H22, 87%, $p < .001$). Not only the number of matching opinions on positions is relevant, but, if another argumentation has only a subset of positions, it can be more important that the priorities are more similar (H21, 74%, $p < .001$).

4.2 Surprising Results

We now have a closer look at more surprising findings from survey which were not in line with the expectations we originally had when designing our hypotheses.

No continuum pro–neutral–contra In Hypothesis 3, we conjectured that a neutral opinion lies exactly between a positive and a negative opinion on a statement. As already mentioned in Section 3, we asked this question in two ways: In variant A, "this cannot be assessed" could be chosen by participants, in variant B, a decision has to be made. In both cases, our expected answer ("neutral" is equally far away from "pro" and "contra") was given by most participants (A: 66%, $p = .006$; B: 95%, $p < 0.001$), where the result is much clearer when forced to make a decision.

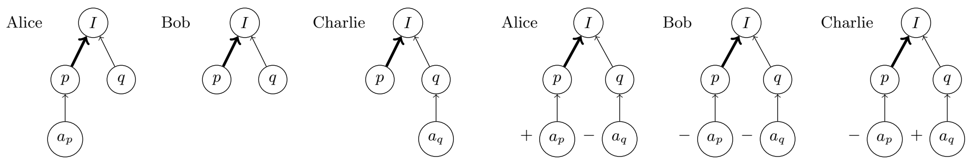
Although the question relevant for us in this scenario was answered as expected, the questions whose attitude is most similar to the positive or negative attitude, respectively, has been answered unexpectedly: We expected that a *positive* opinion is considered closer to *neutral* than to *negative*, but this was only just one of the most frequent answers. In variant B with forced decision, an "equally far apart" assessment has been given by around 50% of the participants.

This can be a hint that many people do not have a mental model where *pro*, *neutral*, and *contra* are arranged in a straight line, but on the corners of a triangle. This might be similar to the opinion triangle presented in [11], with the directions *Belief*, *Disbelief*, and *Ignorance*.

For Hypotheses 7 and 8, we could see similar effects. Hypothesis 7 dealt with whether *no opinion* is equally far away from *pro* and *contra*. For case A, most people give our expected answer (48%, $p = .436$), but many also say that the case cannot be assessed (45%). When forced to make a decision, people choose our expected answer "equally far apart" (95%, $p < 0.001$). But for both variants, we also see the tendency that people have a mental triangle model: In variant B, around 55% have seen *pro* (*contra*) equally far away from *no opinion* and *contra* (*pro*). So being neutral (Hypothesis 3) and having no opinion leads to similar assessments when it is forced, but more people tend to not make an assessment in the *no opinion* case if allowed to.

Lastly, if we consider *pro*, *contra*, and *unknown opinion* (Hypothesis 8), an absolute majority thinks the case cannot be assessed, which makes sense. If a decision is forced, more than 75% percent follow the triangle model again.

Consideration of hierarchies and weights for branches We expected that adding an argument deeper within an argumentation is considered a smaller dissimilarity than adding a new top-level argument (Hypothesis 4, also see Figure 2). This expectation is not confirmed (38%, $p = .36$); the answers are nearly equally distributed across all alternatives. We assume that people count the number of arguments used instead of thinking of an argument hierarchy. Here, further investigations with a more extreme example, e.g. a "deeper" argumentation, would be interesting.



(a) Hypothesis 6 (b) Hypothesis 10

Fig. 4: Visualization of the scenarios for Hypothesis 6 and Hypothesis 10; we expect Bob being closer to Charlie than to Alice in both cases.

Related to this finding are unexpected results for Hypothesis 6: Considering the example depicted in Figure 4a, when comparing Bob with Alice and Charlie, we thought that the similarity to Charlie is greater because the introduced difference is in a branch with lower importance (depicted by a thinner edge). This has not been confirmed, our expected answer is the least frequently chosen answer (24%, $p = .92$). More participants think that Bob is most similar to Alice (40%) or the attitudes are equally far apart (36%).

This is related to the assumption that people do not have a notion for argumentation hierarchy. If people do not catch that a_p and a_q are on the level below p or q, respectively, it makes sense that our expected effect cannot be seen.

But this conjecture is contradicted by the answers for Hypothesis 10, where we thought that not mentioning an argument (as in Figure 4a) and being against an argument (Figure 4b) have the same effect. Our expected answer, Bob is more similar to Charlie than to Alice, is now the most frequently chosen answer (52%, $p = .059$). Thus, our explanations for the unexpected results for Hypothesis 6 do not seem to be correct. Maybe the complexity of the scenario for Hypothesis 10 is so large that people pay closer attention to the nuances of the argumentation. Here, further investigations are necessary.

Trade-off between opinions and arguments Consider a scenario where Alice and Bob have the same opinion on a position, but the arguments are contradictory. Charlie has the same opinions as Alice, but a different opinion on the position. We expect that Alice and Bob are closer than Alice and Charlie (Hypothesis 11) since people probably consider opinions on positions as more important than arguments. Most people answered as we have expected (45%, $p = .174$), but there are also many people saying the attitudes are equally far apart (32%). We can conclude that the common opinion on the position has the greater influence on the assessment of attitude similarity, but arguments also play an important part in the assessment.

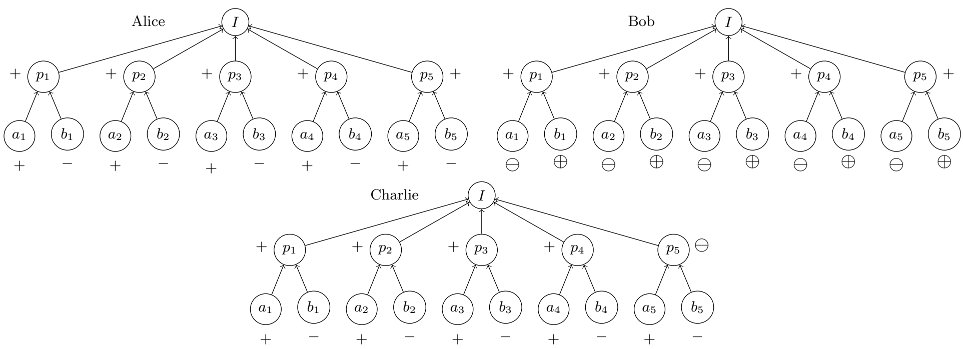


Fig. 5: Visualization of the scenario for Hypothesis 12; differences to Alice are encircled; we expected Alice is considered closer to Charlie than to Bob.

In Hypothesis 12, we assumed not only the opinions on positions are compared, but arguments also play a role and can even “flip” the similarity. For an extreme example with many arguments as shown in Figure 5, our expectation that Alice’s attitude is more similar to Charlie’s has been confirmed (62%,

$p = .004$). This is in line with the findings from Hypothesis 11: Not only common opinions on positions are important for assessing similarity, but also the arguments.

Note that our scenario for Hypothesis 12 converges to the scenario for Hypothesis 11 if p_1 to p_4 are removed. As we have only presented those two extreme scenarios in the questionnaire, we cannot say what the "turning point" is, i.e. what number of common arguments is needed to make up for different opinions.

Opinion tendency vs. weight In Hypotheses 13 and 14, we wanted to know whether an argumentation with e.g. weak positive opinion on a position can be closer to a weak negative opinion on the same position than to a very strong positive opinion. We thought that it is possible, but we were proven wrong. The hypotheses were tested with different formulations and scenarios, as strength/weakness can be expressed in different ways: *strongly for* vs. *slight tendency* (A), *for* vs. *no definite opinion* (B), *strongly for* vs. *doesn't really have an opinion* (C), involving a second, common position (D), and *main reason* vs. *very unimportant reason* (E). Our expected answers were not given by most participants (A: 13% $p = 1.0$; B: 12%, $p = 1.0$; C: 33%, $p = .97$; D: 36%, $p = .95$; E: 34%, $p = .98$), but the similarity to the person with the same direction of opinion has been rated greater (A: 84%, B: 81%, C: 60%, D: 60%, E: 63%).

We can conclude that opinion tendencies are more important than the weights of opinions and arguments.

Alice argues in favor of wind power as follows:

> I am in favor of wind power, as wind turbines do not produce **CO_2 emissions**. Also, I'm for wind power because **wind turbines look nice**.

Bob argues in favor of wind power as follows:

> I am in favor of wind power, as wind turbines do not produce **CO_2 emissions**. I think **wind turbines look nice**, but that is **no argument for wind power** and not relevant for the discussion.

Charlie argues in favor of wind power as follows:

> I am in favor of wind power, as wind turbines do not produce **CO_2 emissions**. I **don't think that wind turbines look nice**.

Fig. 6: Scenario for Hypothesis 15 on the effect of undercuts: We thought that Bob's and Charlie's attitudes are considered equal.

Understanding of undercuts We expected that an opinion belonging to an undercut argument does not count towards the attitude to a position, i.e. in the scenario described in Figure 6, Charlie's and Bob's attitudes are considered equal, regardless whether Charlie's last sentence is mentioned (case A) or not (B). Our results are not clear for this question: "Do Charlie and Alice [or Bob] have the same attitude (opinion and arguments) on wind power?” has been answered with “Yes” by more than 70% in all cases.

We do not understand this result. It could be that the wording of the question for this case is too technical for a good assessment, so that most people only compared the opinions for the position. Another possible explanation is that untrained persons do not understand the undercut attack correctly or find it confusing, and thus fall back to comparing opinions of positions.

Influence of adding new positions in a priority order We wanted to know how the introduction of a new position by a participant influences similarity order. Our anticipation was that it is possible to remove a previous dissimilarity this way (Hypothesis 17), or even swap the similarity order (Hypothesis 18).

Alice:	Bob:	Charlie:	Charlie’:
1. *b*	1. *a*	1. *a*	1. *d*
2. *a*	2. *c*	2. *b*	2. *a*
3. *c*	3. *b*	3. *c*	3. *b*
			4. *c*

(a) Scenario for Hypothesis 17

Alice:	Bob:	Charlie:	Charlie’:
1. *a*	1. *d*	1. *a*	1. *e*
2. *c*	2. *a*	2. *b*	2. *a*
3. *d*	3. *b*	3. *c*	3. *b*
4. *b*	4. *c*	4. *d*	4. *c*
			5. *d*

(b) Scenario for Hypothesis 18

Fig. 7: In these scenarios, Charlie’ introduces a new position not mentioned by the other participants.

To investigate whether those hypotheses can hold, we checked the scenarios depicted in Figure 7. In Figure 7a, we thought that Charlie is considered more similar to Bob (Hypothesis 16), but Charlie’ equally far away from Alice and Bob. The former was confirmed, so changing the order of the most important positions results in a greater perceived difference than flipping less important positions (57%, $p = .018$). The latter was not confirmed (31%, $p = .71$), but we see a clear difference from 57%, indicating that the additional position has an influence on the intuition on similarity. There is no clear “correct” answer, though, since the answers are nearly evenly distributed across all alternatives.

For the scenario in Figure 7b, we anticipated that Charlie is closer to Alice (case A), but Charlie’ closer to Bob (case B; one way to get to this conclusion is counting the number of absolute place differences for each common statement: Charlie–Alice: 4, Charlie–Bob: 6; Charlie’–Alice: 6, Charlie’–Bob: 4). The first expectation has been confirmed (A: 55%, $p = .008$), but not the latter (B: 33%, $p = .64$). In case B, the answers are nearly evenly distributed. Although this is no hint that our hypothesis is sensible, we can see a tendency that the change from case A to B moves the three attitudes closer to each other.

Note that we can neither show that our hypotheses are consistent nor inconsistent, because we only asked for concrete example scenarios. Other scenarios may yield different results, and having results for different scenarios leads to more precise results.

5 Discussion

Our survey was, to our knowledge, the first of its kind. Many results give valuable hints on how an intuitive metric for comparing attitudes expressed in an argumentation should behave. Those metrics have applications in e.g. clustering and recommender systems.

As seen in the previous section, a definite conclusion cannot be drawn for all hypotheses without further surveys. Also, the way we constructed our survey questions could have been suboptimal. We choose a format which is suitable for most Hypothesis to prevent differences due to different formulations of questions. We considered the option to let people rate the similarity of argumentation on a numeric scale, but we thought that this approach is bound to fail: People are unfamiliar with rating argumentation similarity, would probably need some time for "calibration", and the task would feel more unnatural.

Furthermore, the question for "attitude" could have been a problem, because some people may only consider opinions, not arguments. Asking how similar two people "argue" would also be a problem, which we have seen in an internal pretest: Some people started thinking about meta-argumentation aspects, e.g. whether counterarguments are mentioned, or how many arguments are used, and stopped looking at the person's actual attitude.

For questions with ratings of several positions, we switched between complete sentences and enumerations, depending on the number of positions. We thought complete sentences with many positions distract from the actual differences. The change of format could, of course, have an influence, which we did not measure.

We are well aware that MTurk workers are not a representative sample of the US population, and even less for other countries; as already mentioned, the gender distribution does not match the US population. Therefore, generalizing our results for other populations is only possible with caution. Nevertheless, we get some useful insights and hints for further, representative, bigger studies, and possible comparisons between different populations.

6 Related Work

We know no other surveys on attitude similarity in argumentation, but there have been surveys for other purposes to find human baselines.

[14] proposes different measures for determining the similarity of words, and compares the measures with human ratings from a dataset created by [15]. They also think that the quality of a metric can best be determined by comparing it with human common sense. Their dataset contains absolute ratings from 0 (no similarity) to 4 (synonym) for 30 word pairs, each assessed by 38 subjects. We do not think that an absolute rating would have worked for our experiment. First, our argumentation scenarios can have fine-grained or large differences, which probably makes it hard for a person without argumentation theory background to map the difference on a small absolute scale. Second, an absolute scale works well when you can grasp every pair to compare at once and correct older decision

to tweak one's brain scale; this works well with short word pairs, but not with more complex descriptions of argumentation.

In the context of word similarity, [6] find that "comparison with human judgments is the ideal way to evaluate a measure of similarity", which supports our initial assumption that gathering human judgments is important.

In [3], which is based on the study design of [15], 50 human subjects assessed the similarity of process descriptions on a scale from 1 to 5. They compared those assessments with the values of five metrics. Each subject had to indicate how they come to their decision for each comparison, by letting them choose a strategy (e.g. "by process description") from a menu. We did not ask participants how they have come to their assessments. Firstly, we think that reflecting on one's decision influences further decisions. We also think that writing an own description of the decision process is too hard, and providing a menu with possible answers could have influenced following decisions. Moreover, asking this for every question would have significantly increased the length of the questionnaire.

Metrics and applications for comparing argumentations already exist, e.g. based on cosine similarity for opinion prediction [1], and for comparing one's own argumentation with others by counting the number of agreements/disagreements on statements [10]. In both cases, no justification is given why the similarity measure is a good choice. With our work, we want to fill that gap. For instance, we showed that simply counting agreements is not enough.

7 Conclusion and Future Work

We have conducted a survey with human subjects who had to assess the attitude similarity of argumentations. Our results are available for download, and can be used as basis when developing a metric for measuring attitude similarity in argumentation-based applications, e.g. for collaborative filtering. Our results help to transform human gut feeling into a mathematical metric. Some intuitive hypotheses were confirmed by our results, but there were also surprising results, e.g. *neutral* is often not seen as falling on a line between *pro* and *con*.

Our survey cannot establish "absolute truths", but we have collected first hints on what properties a metric which matches human intuition should have. In future work, we want to compare several metrics to see which properties they fulfill and how that matches human intuition. Moreover, further research is needed for hypotheses where we could not get clear results, and where there are turning points in trade-off scenarios. Also, more representative surveys and a comparison of different countries are needed.

References

1. Althuniyan, N., Sirrianni, J.W., Rahman, M.M., Liu, X.F.: Design of mobile service of intelligent large-scale cyber argumentation for analysis and prediction of collective opinions. In: International Conference on AI and Mobile Services. pp. 135–149. Springer (2019)

2. Berinsky, A.J., Huber, G.A., Lenz, G.S.: Using mechanical turk as a subject recruitment tool for experimental research (2011)
3. Bernstein, A., Kaufmann, E., Bürki, C., Klein, M.: How similar is it? towards personalized similarity measures in ontologies. In: Wirtschaftsinformatik 2005, pp. 1347–1366. Springer (2005)
4. Brenneis, M., Behrendt, M., Harmeling, S., Mauve, M.: How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs. In: Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020). pp. 2–13. No. 2672 in CEUR Workshop Proceedings, Aachen (Sep 2020)
5. Brenneis, M., Mauve, M.: deliberate – Online Argumentation with Collaborative Filtering. In: Computational Models of Argument. vol. 326, p. 453–454. IOS Press (Sep 2020)
6. Budanitsky, A., Hirst, G.: Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and other lexical resources. vol. 2, pp. 2–2 (2001)
7. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (1995)
8. Ebbinghaus, B., Mauve, M.: decide: Supporting Participatory Budgeting with Online Argumentation. In: Computational Models of Argument. Proceedings of COMMA 2020. Frontiers in Artificial Intelligence and Applications, vol. 326, p. 463–464. IOS Press (Sep 2020)
9. Franklin, C.H.: The 'margin of error' for differences in polls. See https://abcnews.go.com/images/PollingUnit/MOEFranklin.pdf (2007)
10. Gordon, T.F.: Structured consultation with argument graphs. From Knowledge Representation to Argumentation in AI. A Festschrift in Honour of Trevor Bench-Capon on the Occasion of his 60th Birthday pp. 115–133 (2013)
11. Haenni, R.: Probabilistic argumentation. Journal of Applied Logic **7**(2), 155–176 (2009)
12. Krauthoff, T., Meter, C., Mauve, M.: Dialog-Based Online Argumentation: Findings from a Field Experiment. In: Proceedings of the 1st Workshop on Advances in Argumentation in Artificial Intelligence. pp. 85–99 (November 2017)
13. Kunz, W., Rittel, H.W.J.: Issues as elements of information systems, vol. 131. Citeseer (1970)
14. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on knowledge and data engineering **15**(4), 871–882 (2003)
15. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and cognitive processes **6**(1), 1–28 (1991)
16. Neugebauer, D.: DABASCO: Generating AF, ADF, and ASPIC+ Instances from Real-World Discussions. In: Computational Models of Argument. Proceedings of COMMA 2018. pp. 469–470 (2018)
17. Scott, A.J., Seber, G.A.: Difference of proportions from the same survey. The American Statistician **37**(4a), 319–320 (1983)
18. Silvapulle, M.J., Sen, P.K.: Constrained Statistical Inference: Order, Inequality, and Shape Constraints, vol. 912. John Wiley & Sons (2011)

3.3.3 Hypothesis Test for the Difference of Two Multinomial Proportions From the Same Sample

To evaluate the results of the survey, we asked ourselves how significant the results we got were. As we know of no thorough description of a statistical test relevant for our case, we give a general description of the method we used. Afterwards, we proof the requirements for the theorem used in the test procedure and then have a look at a small simplification.

Derivation of the Test Procedure

Consider we want to do a survey to find out the favorite color of a group of people. How can we check if our conjecture is backed by the survey results reasonably well?

As an example, let us conjecture that blue is the most popular (relative frequency) primary color in Germany. We survey $n = 100$ randomly chosen Germans, who have to give exactly one answer. The answer options and their answer frequencies are:

1. Blue: 45
2. Yellow: 35
3. Red: 20

We want to find out whether blue is just "randomly" the most frequently chosen answer (i.e. if blue just randomly has more votes than yellow and red and is not the most popular color). Our null hypothesis H_0 is: "Blue is not more popular than yellow, or blue is not more popular than red." It is a disjunction of sub-hypotheses H_{01} "Blue is not more popular than yellow." and H_{02} "Blue is not more popular than red." Our goal is to calculate the p-value, which is an estimate for the probability of getting these extreme (or even more extreme) results if H_0 is true. We will reject H_0 if the p-value is less than a significance level α, e.g. $\alpha = 0.1$.[1]

The following requirements apply:

1. n is considerably smaller than the total population N ($n/N < 0.05$) in order to approximate the hypergeometric distribution with the binomial distribution.
2. The sample is randomly drawn from the total population and is representative.

Let us first look at how we calculate a p-value for H_{01}. Our problem of determining whether "blue" received significantly more votes than any other answer is analogous to the problem of determining whether the winner of an election received significantly more votes, which was addressed by Franklin (2007).

[1] If we talk about a level α, we are *not* referring to the parameter α of our pseudometric.

Let $P_B = 0.45$ and $P_Y = 0.35$ be the relative frequencies of the two options examined. The variance for the difference $P_B - P_Y$ is (based on the difference of two multinomial proportions (Scott and Seber, 1983))

$$\texttt{Var}(P_B - P_Y) = \frac{(P_B + P_Y) - (P_B - P_Y)^2}{n}. \tag{3.22}$$

Using the values of our example, a one-sided Z-test yields the following result for H_{01}:

$$\texttt{Var}(P_B - P_Y) = 0.0079 \tag{3.23}$$

$$z_1 = \frac{P_B - P_Y}{\sqrt{\texttt{Var}(P_B - P_Y)}} = 1.125 \tag{3.24}$$

$$p_1 = \Phi(-z) = 0.130 \tag{3.25}$$

Analogously, we get $p_2 = 0.0006$ for H_{02}.

With an *intersection–union test* (IUT) (Silvapulle and Sen, 2011, p. 240) and by applying Theorem 2 of Berger and Hsu (1996), we get $p = \max(p_1, p_2) = 0.130$ for H_0.

As $p > \alpha$, there is no strong enough indication that H_0 should not be correct. On the basis of the survey results, we cannot reject with sufficient certainty our null hypothesis that blue is not the most popular primary color in Germany.

Note that if there is no prior hypothesis about which or whether an answer is chosen most frequently, it can be tested first whether there is a most frequent answer (e.g. with a χ^2 test), and then whether the most frequent answer is given significantly more often. However, to maintain the significance level, corrections must be made for testing in the same sample (keyword *family-wise error rate* (FWER), for example, using the Holm–Bonferroni method (Holm, 1979)).

Proof for the Requirements of Theorem 2

We use the following theorem above:

Theorem 1 (Theorem 2 in (Berger and Hsu, 1996)) *For some $i = 1, \ldots, k$, suppose R_i is a size-α rejection region for testing H_{0i} vs. H_{Ai}. For every $j = 1, \ldots, k$, $j \neq i$, suppose R_j is a level-α rejection region for testing H_{0j} vs. H_{Aj}. Suppose there exists a sequence of parameter points θ_l, $l = 1, 2, \ldots,$ in Θ_i such that*

$$\lim_{l \to \infty} P_{\theta_l}(\mathbf{X} \in R_i) = \alpha,$$

and, for every $j = 1, \ldots, k$, $j \neq i$,

$$\lim_{l \to \infty} P_{\theta_l}(\mathbf{X} \in R_j) = 1.$$

Then the intersection–union test with rejection region $R = \bigcap_{i=1}^{k} R_i$ is a size-α test of H_0 vs. H_A.

It still has to be shown that the requirements of the theorem are fulfilled. Without loss of generality, let us consider the case that we have three answer options. The test for the sub-hypotheses is of size-α by design. For (P_B, P_Y, P_R, n), we look at the series $(0.5 + \frac{\Phi^{-1}(-\alpha)}{2\sqrt{(\Phi^{-1}(-\alpha))^2+n}}, 0.5 - \frac{\Phi^{-1}(-\alpha)}{2\sqrt{(\Phi^{-1}(-\alpha))^2+n}}, 0, n)_{n\in\mathbb{N}}$. We get the probability of *not* being in the rejection area:

$$
\begin{aligned}
\lim_{n\to\infty} \Phi\left(-\frac{P_B - P_Y}{\sqrt{\frac{(P_B+P_Y)-(P_B-P_Y)^2}{n}}}\right) &= \lim_{n\to\infty} \Phi\left(-\frac{\frac{\Phi^{-1}(-\alpha)}{\sqrt{(\Phi^{-1}(-\alpha))^2+n}}}{\sqrt{\frac{1-\frac{(\Phi^{-1}(-\alpha))^2}{(\Phi^{-1}(-\alpha))^2+n}}{n}}}\right) && (3.26)\\
&= \lim_{n\to\infty} \Phi\left(-\frac{\frac{\Phi^{-1}(-\alpha)}{\sqrt{(\Phi^{-1}(-\alpha))^2+n}}}{\sqrt{\frac{\frac{n}{(\Phi^{-1}(-\alpha))^2+n}}{n}}}\right) && (3.27)\\
&= \lim_{n\to\infty} \Phi\left(-\frac{\frac{\Phi^{-1}(-\alpha)}{\sqrt{(\Phi^{-1}(-\alpha))^2+n}}}{\sqrt{\frac{1}{(\Phi^{-1}(-\alpha))^2+n}}}\right) && (3.28)\\
&= \lim_{n\to\infty} \Phi(-\Phi^{-1}(-\alpha)) && (3.29)\\
&= 1-\alpha && (3.30)
\end{aligned}
$$

and

$$
\begin{aligned}
\lim_{n\to\infty} \Phi\left(-\frac{P_B - P_R}{\sqrt{\frac{(P_B+P_R)-(P_B-P_R)^2}{n}}}\right) &= \lim_{n\to\infty} \Phi\left(-\frac{P_B}{\sqrt{\frac{P_B-P_B^2}{n}}}\right) && (3.31)\\
&= 0 && (3.32)
\end{aligned}
$$

For Equation (3.31), remember that we chose $P_R = 0$.

Consideration of the Most Frequent Answers Sufficient

It is sufficient to look only at P_B and the relative frequency of the most frequently chosen option that is not "blue" (P_Y in our case) to calculate the p-value for the IUT. This is true, because, given $P_Y \geq P_R$:

$$\begin{aligned}
p_1 &= \Phi(-z) && (3.33)\\
&= \Phi\left(-\frac{P_B - P_Y}{\sqrt{\mathtt{Var}(P_B - P_Y)}}\right) && (3.34)\\
&= \Phi\left(-\frac{P_B - P_Y}{\sqrt{\frac{(P_B+P_Y)-(P_B-P_Y)^2}{n}}}\right) && (3.35)\\
&\geq \Phi\left(-\frac{P_B - P_R}{\sqrt{\frac{(P_B+P_R)-(P_B-P_R)^2}{n}}}\right) && (3.36)\\
&= p_2 && (3.37)
\end{aligned}$$

For (3.35) $\geq$ (3.36), remember that the cumulative distribution function of the normal distribution Φ is monotonically increasing.

3.3.4 Impact on Our Desiderata and Limitations

As we now know what properties can be regarded "intuitive" with high certainty, we can use our results from Subsection 3.3.2 to check whether our desiderata from Subsection 3.2.2 are actually intuitive or not.

Desiderata 1 (*Proportionally bigger overlap is better*) and 5 (*Weights of arguments have influence even if they are the only difference*) were checked in Hypotheses 1 and 5, respectively, and can be considered confirmed intuitive.

Desideratum 2 (*Contrary opinion is worse than no opinion*) could not be confirmed as expected: We checked it in different scenarios: with no, neutral, and unknown opinion, and, for each case, with the possibility to say that the case cannot be assessed (Hypotheses 3, 7 and 8). We could see that many people, especially for no and unknown opinion, seemed to create a mental triangle, so that *no/unknown opinion*, *pro*, and *contra* had the same pairwise distance.

Our question for Desideratum 3 (*Deviation in deeper parts has less influence than deviation in higher parts*, Hypothesis 4) did not get a clear winner, so we can neither confirm our original expectation, nor do we know what the actual "truth" is for this question.

Desideratum 4 (*Influence of deeper parts depends on weights in higher parts*) was checked with Hypotheses 6 and 10. Here we got some inconsistent results and further research is needed.

The trade-offs between argument weights and agreement (Desideratum 7, Hypothesis 14), and statement ratings and agreement (Desideratum 8, Hypothesis 13) could not be shown in the expected form. In contrast to what we originally expected, people assessed the influence of the side of the opinion much greater in comparison to the weight of the opinion. Note, however, that the consequences of our wrong conjecture can be mitigated by improving the mapping of user input on the rating values, e.g. by not allowing ratings near 0.

We have also mentioned possible limitations of the proposed pseudometric. With the help of the survey, we could check whether these limitations are actually against intuition.

The first limitation was about ignoring everything underneath an edge with weight 0. This motivated Hypothesis 15, for which, unfortunately, no good results were achieved, as mentioned in the paper (Brenneis and Mauve, 2020b): When comparing the attitude in all three cases of the scenario, people completely ignored all arguments. Thus, we cannot say from our survey whether this limitation is an "intuitive" limitation.

The second limitation was related to changing the order by adding a new, unrelated opinion on a new position, which corresponds to Hypothesis 18. For the relevant question whether that complete change is intuitive, no significant answer could be found, i.e. we could not show that not having that change is unintuitive. But we were able to see that adding another opinion can lead to not having a clear right answer. Hence, as untrained people did not agree on a single correct answer, it is okay for a metric to choose any answer as "correct." It is, though, sensible to research further with similar scenarios.

3.4 Comparing Different Distance Functions Against a Human Baseline

We now have a baseline about how average humans assess the similarity of certain argumentation scenarios, but we do not yet know how well our pseudometric from Section 3.2 matches this baseline. Maybe we can find other distance functions which are even more in line with human intuition.

In this section, we first introduce several other possible distance functions based on functions which can be found in related work. Then we expound on the functions' performances when compared with our human baseline by looking at our publication by Brenneis and Mauve (2021c). To wrap up, we examine some more insights we gain from our human baseline, especially with regard to our pseudometric. In the end, we will see that, depending on the argumentation structure of the application context, simpler distance functions can be more suitable than our quite complex pseudometric.

3.4.1 Introduction to the Different Distance Functions

In Subsection 3.4.2, we will compare seven different distance functions which are supposed to compare attitudes in argumentations:

1. Tree similarity measure by Bhavsar et al. (2004)

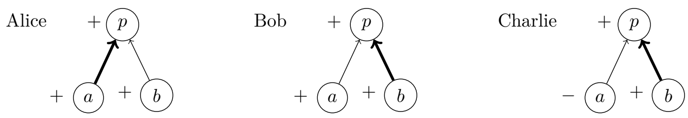


Figure 3.2: Example graphs for the calculation examples (based on Hypothesis 14).

2. Cosine distance (similar to the work by Rahman et al. (2019))
3. Jaccard distance (used by Kunze et al. (2011))
4. p-metric for fuzzy sets (Xuecheng, 1992)
5. Soergel distance (also used by Rahman et al. (2019))
6. VAA distance (as used in several VAAs with proximity voting logic (Romero Moreno et al., 2020))
7. our pseudometric (Brenneis et al., 2020)

To get a better feeling of how the different distance functions work before doing the actual comparison, we will now examine some calculation examples. Each distance function is formally defined in Subsection 3.4.2, possibly more generally with different parametrizations. We calculate the distance between Alice's and Bob's, as well as Alice's and Charlie's graph in Figure 3.2. For the calculation examples, we use simple parametrizations of the functions to get an idea of the intuition behind every distance function.

The examples of the scenarios in the comparison do not make use of all the features which can be represented in a weighted argumentation graph (defined in Subsection 3.2.2). To simplify the notation of the distance functions, two simpler kinds of representations are used for the graphs. One possibility is a representation as a pair of two sets, where the first set captures the opinions on statements, the second set the weights of edges:

$$A = (\{(p, \texttt{+}), (a, \texttt{+}), (b, \texttt{+})\}, \{((p, I), 1), ((a, p), 1), ((b, p), 2)\}) \tag{3.38}$$
$$B = (\{(p, \texttt{+}), (a, \texttt{+}), (b, \texttt{+})\}, \{((p, I), 1), ((a, p), 2), ((b, p), 1)\}) \tag{3.39}$$
$$C = (\{(p, \texttt{+}), (a, \texttt{-}), (b, \texttt{+})\}, \{((p, I), 1), ((a, p), 2), ((b, p), 1)\}) \tag{3.40}$$

Here, `+` stands for a strong agreement, `-` for a strong disagreement, and `(+)` would be a weak agreement. The weights are translated to a priority order, which means 1 is a thick edge (high weight), 2 a thin edge (low weight).

The other representation used by some distance functions is a vector representation. Each vector captures how much each statement is agreed to and how much each statement is disagreed to, each on a range from 0 to 0.5. In this example, we hence get six vector components which capture the opinions in the order "agreement with p," "disagreement with p," "agreement with a," and so on, plus three components capturing the weight of the arguments. Our integer edge weights can be mapped differently to weights from the interval $[0, 1]$. For the example calculations, we use a normalized harmonic mapping (as defined in Subsection 3.4.2), where the bold edges in our example get a weight of $\frac{\frac{1}{1}}{\frac{1}{1}+\frac{1}{2}} = \frac{2}{3}$, and thin edges a weight of $\frac{\frac{1}{2}}{\frac{1}{1}+\frac{1}{2}} = \frac{1}{3}$. We get the following vectors for the graphs in Figure 3.2:

$$A = \left(0.5, 0, 0.5, 0, 0.5, 0, 1, \frac{2}{3}, \frac{1}{3}\right) \tag{3.41}$$

$$B = \left(0.5, 0, 0.5, 0, 0.5, 0, 1, \frac{1}{3}, \frac{2}{3}\right) \tag{3.42}$$

$$C = \left(0.5, 0, 0, 0.5, 0.5, 0, 1, \frac{1}{3}, \frac{2}{3}\right) \tag{3.43}$$

where we write edge weights as fractions to discern them from values concerning statement opinions.

With these two kinds of mathematical representation, we now examine how the seven distance functions calculate the dissimilarity of the graphs.

The tree similarity measure by Bhavsar et al. (2004) is lengthy to describe, but simplifies to the following for this example: If node labels (i.e. opinions) are all the same and the weights of edges are the only difference, the difference of the trees is 0. A similarity bonus N is awarded if an intermediate node has common labels (like p in our example). The N serves a similar purpose as α in our own distance function. We use $N = 0.5$ for this example. If two node labels are the same, the similarity is 1, and weighted by $1 - N$ and the averaged edge weights from both trees on the path from the root I to the node; if the labels are different, the similarity is 0. Hence, we get the following distances (using normalized harmonic transformation for the weights):

$$d(A, B) = 0 \tag{3.44}$$

$$d(A, C) = 1 - \left(0.5 + (1 - 0.5) \cdot (1 + (0.5 - 0.5) \cdot \left(\frac{\frac{2}{3} + \frac{1}{3}}{2} \cdot 0 + \frac{\frac{1}{3} + \frac{2}{3}}{2} \cdot 1\right)\right) = 0.125 \tag{3.45}$$

For the Cosine distance, the graphs are transformed to vectors, and we then calculate the distances as:

$$d(A, B) = 1 - \frac{A \cdot B}{||A||\,||B||} = 1 - \frac{3 \cdot 0.5^2 + 1^2 + \frac{2}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3}}{0.5^2 \cdot 3 + 1^2 + \frac{2^2}{3^2} + \frac{1}{3^2}} \approx 0.0482 \tag{3.46}$$

$$d(A, C) = 1 - \frac{A \cdot C}{||A||\,||C||} = 1 - \frac{2 \cdot 0.5^2 + 1^2 + \frac{2}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3}}{0.5^2 \cdot 3 + 1^2 + \frac{2^2}{3^2} + \frac{1}{3^2}} \approx 0.157 \tag{3.47}$$

The Jaccard distance is applied on the set representation, where both sets of the tuples are united. We abbreviate those sets as

$$A_\cup = \{(p, +), (a, +), (b, +), ((p, I), 1), ((a, p), 1), ((b, p), 2)\} \tag{3.48}$$

$$B_\cup = \{(p, +), (a, +), (b, +), ((p, I), 1), ((a, p), 2), ((b, p), 1)\} \tag{3.49}$$

$$C_\cup = \{(p, +), (a, -), (b, +), ((p, I), 1), ((a, p), 2), ((b, p), 1)\} \tag{3.50}$$

The symmetric difference counts the number of different items in the set, i.e., in principle, the number of different opinions and weights is counted.

$$d(A,B) = \frac{|A_\cup \triangle B_\cup|}{|A_\cup \cup B_\cup|} = \frac{4}{8} = \frac{1}{2} \tag{3.51}$$

$$d(A,C) = \frac{|A_\cup \triangle C_\cup|}{|A_\cup \cup C_\cup|} = \frac{6}{9} = \frac{2}{3} \tag{3.52}$$

Our variant of the p-metric calculates the distance between opinions, plus the distance of the argument weight. For the weights, we take a normalized harmonic transformation here again, and we choose $p = 2$:

$$d(A,B) = \left(|0.5-0.5|^2 + |0.5-0.5|^2 + |0.5-0.5|^2 + |1-1|^2 + \left|\frac{2}{3}-\frac{1}{3}\right|^2 + \left|\frac{1}{3}-\frac{2}{3}\right|^2\right)^{\frac{1}{2}} \tag{3.53}$$

$$\approx 0.471 \tag{3.54}$$

$$d(A,C) = \left(|0.5-0.5|^2 + |0.5-(-0.5)|^2 + |0.5-0.5|^2 + |1-1|^2 + \left|\frac{2}{3}-\frac{1}{3}\right|^2 + \left|\frac{1}{3}-\frac{2}{3}\right|^2\right)^{\frac{1}{2}} \tag{3.55}$$

$$\approx 1.11 \tag{3.56}$$

The Soergel distance is also applied to the vector representation. For each vector component, the numerator contains the minimum of a component's value from both input vectors, the denominator the maximum. This distance function can be regarded as a weighted variant of the Jaccard distance.

$$d(A,B) = 1 - \frac{0.5+0+0.5+0+0.5+0+1+\frac{1}{3}+\frac{1}{3}}{0.5+0+0.5+0+0.5+0+1+\frac{2}{3}+\frac{2}{3}} \approx 0.174 \tag{3.57}$$

$$d(A,C) = 1 - \frac{0.5+0+0+0+0.5+0+1+\frac{1}{3}+\frac{1}{3}}{0.5+0+0.5+0.5+0.5+0+1+\frac{1}{3}+\frac{1}{3}} \approx 0.385 \tag{3.58}$$

We also look at how a VAA would calculate the distance. As VAAs only look at positions, and the opinion and rating of the position is the same in all three graphs, we get a distance of 0 for both cases:

$$d(A,B) = 2 \cdot 1 \cdot |0.5 - 0.5| = 0 \tag{3.59}$$

$$d(A,C) = 2 \cdot 1 \cdot |0.5 - 0.5| = 0 \tag{3.60}$$

The factor 2 is introduced because the opinions in the graphs are *strong*, which corresponds to double weighted positions in a VAA.

Last but not least, we considered our own pseudometric, which we have already explained in detail in Section 3.2. With $\alpha = 0.5$ and normalized harmonic weight transformation, we get:

$$d(A,B) = (1-0.5)\left(0.5^1 \cdot |0.5-0.5| + 0.5^2 \cdot \left|\frac{2}{3}\cdot 0.5 - \frac{1}{3}\cdot 0.5\right| + 0.5^2 \cdot \left|\frac{1}{3}\cdot 0.5 - \frac{2}{3}\cdot 0.5\right|\right) \tag{3.61}$$

$$\approx 0.0417 \tag{3.62}$$

$$d(A,B) = (1-0.5)\left(0.5^1 \cdot |0.5-0.5| + 0.5^2 \cdot \left|\frac{2}{3}\cdot 0.5 - \frac{1}{3}\cdot (-0.5)\right| + 0.5^2 \cdot \left|\frac{1}{3}\cdot 0.5 - \frac{2}{3}\cdot 0.5\right|\right) \tag{3.63}$$

$$\approx 0.0833 \tag{3.64}$$

From the calculation examples, one can see that different distance functions consider different sets of information. For instance, some functions do not consider how deep a statement is within an argumentation graph. Note, however, that the similarity order created by the majority of distance functions is the same in this example; the only exception is the VAA distance. We will now examine the implications of those differences on the "intuitiveness" of the functions.

3.4.2 Comparison of the Distance Function With Our Human Baseline

After having understood how the distance functions basically work, we now formally define how they calculate the distance between two argumentation graphs and compare the calculated distances with our human baseline results. For this purpose, we regard the paper by Brenneis and Mauve (2021c):

Markus Brenneis and Martin Mauve.

"How Intuitive Is It? Comparing Metrics for Attitudes in Argumentation with a Human Baseline"

In: *Proceedings of the 2nd International Conference on Artificial Intelligence in HCI, held as part of HCI International 2021*, *Artificial Intelligence in HCI*, pages 125–138, Springer International Publishing.
Acceptance Rate: 58.07%

Within this publication, we made the following contributions:

1. adaption of six distance functions for comparing attitudes in argumentations
2. comparison of seven distance functions with our human baseline
3. interpretation why certain distance functions are better than others in regard to certain properties

Note that only an Extended Abstract has been peer-reviewed. For completeness, we include both, the peer-reviewed Extended Abstract and the published full paper, in this section.

Personal Contribution

Markus Brenneis, the author of this thesis, had the idea for comparing different distance functions for assessment of argumentation similarity. He conducted the original survey this paper is based on (presented in Section 3.3), adapted the distance functions to a common argumentation model, implemented the different distance functions to evaluate them, and wrote the whole paper. Martin Mauve provided feedback on drafts of the paper, especially regarding the structure.

Importance and Impact on This Thesis

With this work, we checked whether our pseudometric presented in Section 3.2 actually matches human intuition and thus yields results which can be trusted. The baseline for human intuition has been collected before and was described in Section 3.3. The distance function which matches human intuition best would be most suitable for use in applications which need such a function, e.g. our VAA we present in Section 4.2. Those applications need distance results which are understandable, intuitive, and do not feel “random.”

Extended Abstract – How Intuitive Is It? Comparing Metrics for Attitudes in Argumentation with a Human Baseline

Markus Brenneis and Martin Mauve

Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany
Markus.Brenneis@uni-duesseldorf.de

1 Introduction

Comparing attitudes different people or organizations have in an argumentation is often relevant and useful, e.g. for clustering using opinions mentioned in argumentations, finding a consensus, recommender systems for argumentation platforms (such as our platform *deliberate* [3], which can be used for political education), or comparing one's own attitudes and arguments with those of political parties. Especially if used for sensitive tasks like recommending a party to vote for, it is important to have a distance measure which yields intuitive results which can be understood. In previous work [4], we have conducted a survey with untrained human subjects to find out what properties a distance function for argumentation data should fulfill to yield results matching human intuition.

In this work, we compare different distance functions regarding those properties. Our goal is to provide hints for applications which kinds of distance functions best match human intuition and where and why there are differences. In our argumentation model we consider that arguments can be of different strengths, and persons can be more or less sure about their opinions, which should be considered when calculating the distance between persons.

Our contribution is the following: First, we present a list of properties which should be fulfilled by a distance function which compares argumentations, based on a survey we have conducted earlier. We adapted different existing distance functions to use them with attitudes in argumentations. Then we compare those functions regarding different properties we found to be intuitive through our survey, and examine different values for the hyperparameters of each function. Finally, we discuss why the distance functions fail to fulfill some properties.

2 Comparison of Distance Functions for Argumentations

Seven distance functions are included in our comparison, of which most are based on previous works in argumentation theory or related fields, and which have been adapted by us for use with our formal definition of argumentation graphs which consider strengths of arguments and statements [2]. Many functions have different hyperparameters, for which we tested different values. We included the following functions:

- Bhavsar distance [1] (originally used for match-making of agents in e-business environments)
- Cosine distance (similar to [6], who predict opinions in argumentation)
- Jaccard distance (used in [5] as basis for calculating the similarity of process models)
- p-metric for fuzzy sets [8]
- Soergel distance (also used by [6])
- VAA distance (as used in different Voting-Advice Applications with proximity voting logic [7])
- our weighted argumentation tree distance (WATD) [2]

We think the best way to check whether a distance function is intuitive is comparing it with a human baseline. In an online survey we have previously conducted[1] [4], different possible properties for distance functions comparing attitudes in argumentation settings have been checked for their intuitiveness. Assessments by untrained human subjects have been collected for different argumentation scenarios. From the survey results, we got a list of properties which should be fulfilled by a distance function to match human intuition. If we look only at properties which can be considered intuitive from that survey on a significance level $\alpha = 10\%$, we get a list of 22 properties which should be fulfilled, i.a.

1. weights of arguments have an influence even if they are the only difference,
2. *no opinion* has the same distance from a *positive* and a *negative* opinion,
3. flipping the order of two most important positions results in a bigger difference than flipping two less important positions,
4. the distance between an unknown opinion and a positive (or negative) opinion is the same as between a positive and a negative opinion.

Most properties are fulfilled by the p-metric (21 properties), Cosine, and Soergel distance (20); VAA has the worst result (8). The VAA distance does badly because it cannot deal with small weight differences and does not consider deeper arguments.

Some functions fail with some properties by design, e.g. the Bhavsar distance explicitly ignores weights if they are the only difference [1, Example 2], contradicting properties 1 and 3. Property 4 is only fulfilled by the p-metric and the Jaccard distance; the former explicitly defines every comparison with an unknown opinion as 1, the latter treats any difference of opinion as equally distant. Other functions, e.g. WATD, are defined to treat an unknown opinion as falling between positive and negative opinion, which does not match the intuition of average human subjects.

From our evaluation, one gets an idea which metrics yield intuitive results for applications which compare attitudes in argumentations, although our approach has some limitations. For instance, we had no look at bigger argumentation hierarchies, as our previous survey did not give significant results for them. Thus, further research is needed.

[1] raw data at `https://github.com/hhucn/argumentation-similarity-survey-results/`

References

1. Bhavsar, V.C., Boley, H., Yang, L.: A weighted-tree similarity algorithm for multi-agent systems in e-business environments. Computational Intelligence **20**(4), 584–602 (2004)
2. Brenneis, M., Behrendt, M., Harmeling, S., Mauve, M.: How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs. In: Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020). pp. 2–13. No. 2672 in CEUR Workshop Proceedings, Aachen (Sep 2020)
3. Brenneis, M., Mauve, M.: deliberate – Online Argumentation with Collaborative Filtering. In: Computational Models of Argument. vol. 326, p. 453–454. IOS Press (Sep 2020). https://doi.org/10.3233/FAIA200530
4. Brenneis, M., Mauve, M.: Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations. In: Proceedings of the Workshop on Advances In Argumentation In Artificial Intelligence 2020. pp. 1–15. No. 2777 in CEUR Workshop Proceedings, Aachen (Nov 2020)
5. Kunze, M., Weidlich, M., Weske, M.: Behavioral similarity – a proper metric. In: International Conference on Business Process Management. pp. 166–181. Springer (2011)
6. Rahman, M.M., Sirrianni, J., Liu, X.F., Adams, D.: Predicting opinions across multiple issues in large scale cyber argumentation using collaborative filtering and viewpoint correlation. The Ninth International Conference on Social Media Technologies, Communication, and Informatics pp. 45–51 (2019)
7. Romero Moreno, G., Padilla, J., Chueca, E.: Learning VAA: A new method for matching users to parties in voting advice applications. Journal of Elections, Public Opinion and Parties pp. 1–19 (2020)
8. Xuecheng, L.: Entropy, distance measure and similarity measure of fuzzy sets and their relations. Fuzzy sets and systems **52**(3), 305–318 (1992)

How Intuitive Is It? Comparing Metrics for Attitudes in Argumentation with a Human Baseline*

Markus Brenneis and Martin Mauve

Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany
Markus.Brenneis@uni-duesseldorf.de

Abstract. It is often interesting to know how similar two persons argue, e.g. when comparing the attitudes of voters and political parties, or when building an argumentation-based recommender system. Those applications need a distance function, which should give intuitive results. In this paper, we present seven functions which calculate how similar the attitudes of two agents are in an argumentation. We evaluate how good those functions match the results of a human baseline which we determined in a previous work. As it turns out, variants of the p-metric, Cosine, and Soergel distance best agree with human intuition.

Keywords: Argumentation · Metric · Human Baseline.

1 Introduction

Comparing the attitudes different people or organizations have in an argumentation is often relevant and useful, e.g. for clustering using opinions mentioned in argumentations, recommender systems for argumentation platforms (as used in our platform *deliberate* [4]), or comparing one's own attitudes and arguments with those of political parties. In a previous work [5], we have conducted a survey with untrained human subjects to find out what properties a distance function for argumentation data should fulfill to yield results matching human intuition.

In this paper, we compare different distance functions regarding those properties. Our goal is to provide hints for application developers which kinds of distance functions best match human intuition and where and why there are differences. This helps with choosing functions best suited for the problem at hand, knowing that their results follow intuitive and understandable properties.

Our contribution is the following: We present a list of properties which should be fulfilled by a distance function which compares argumentations, based on a survey we have conducted earlier. Different existing distance functions were

* Manchot research group *Decision-making with the help of Artificial Intelligence*, use case politics
The final authenticated version is available online at https://doi.org/10.1007/978-3-030-77772-2_9.

adapted to use them with attitudes in argumentations. We compare those functions regarding different properties we found to be intuitive through our survey, and examine different values for the hyperparameters of each function. Afterwards, we explain why certain functions perform better than others.

In the next section, we provide the key definitions used throughout our work. Afterwards, we define the formal mathematical model and distance functions we compared. We then present and discuss the results of comparing the functions with a human baseline, and finally have a look at related work.

2 Definitions

The argumentative terms we use in this paper are based on the IBIS model [10] for argumentation. An argumentation consists of *arguments*, and each argument is formed by two *statements*: a *premise* and a *conclusion*. We call the set of all statements S and the set of all arguments $A \subset S^2$.

A special "statement" is the *issue* I, which denotes the topic of an argumentation and has no conclusions. All premises for arguments with I as the conclusion are referred to as *positions*, and are typically actionable items like "We should build more wind power plants." $P \subset S$ is the set of all positions.

Different persons can have individual views in an argumentation: They can (strongly) agree (denoted as `(+)` (agree), or `+` (strongly agree), respectively) or disagree (`(-)`, `-`) with statements[1], be neutral (`0`) about a statement, indicate to not have an opinion (`∅`), or do not mention anything about a statement (`?`; so we do not know their opinion); we call this stance on statements *opinion*. We define the set of possible opinion values for a statement $O := \{\texttt{+}, \texttt{-}, \texttt{(+)}, \texttt{(-)}, \texttt{0}, \emptyset, \texttt{?}\}$.

They can also assign arguments different *relevances* (or *weights* or *importances*), and give a priority order for positions. The overall importances and opinions of a person are referred to as their *attitude*.

We represent a person's attitude as an argumentation tree[2], or, if only positions are involved, as sorted lists with positions, where the most important position is at the top. Note that in our tree representation, statements are nodes, argument are edges, to have statements as atomic building blocks. This visualization can, however, be transformed to classical Dung-based [7] abstract argumentation frameworks when needed. We do not draw the common root I in our visualizations to make them simpler.

As an example, we explain how Alice's tree in Figure 1e should be understood: Alice agrees with the position p and the statements a and b, which build arguments with conclusion p. The argument (a, p) is more important for her than the argument (b, p) (indicated by the bolder edge). Note that we do not differentiate whether an argument edge is attacking or defending – this is up to the interpretation of the natural language presentation of the scenario, but

[1] A more fine-grained model for the strength of (dis-)agreement, as we have suggested in [3], could be used, but is not necessary in this work.

[2] A representation as more general graphs is also possible, but again not necessary for the examples in this work.

is consistent within all trees of one scenario (i.e. in Figure 1e, the edges (a, p) in all three trees are either consistently attacking or supporting arguments); a differentiation is therefore not needed in the model for the purpose of this paper.

Throughout this paper, we use the term *distance function* to refer to a function which calculates some distance between pairs of argumentations with the parts introduced above. Those functions might happen to fulfill all properties of a metric (e.g. the triangle equality), but are not required to do so.

We now define how the drawing of a tree is translated to mathematical objects. Each tree can be considered as a pair of functions (o, s), where $o : S \to O$ captures the opinion on statements, $s : A \to \mathbb{N}_0$ the sorting of arguments by importance (where 1 means top-priority, 0 no priority (as default for not mentioned arguments); the ordering is not required to be injective). Note that we view a function as a set of ordered pairs (parameter, function value).

Please note the following conventions: The sorting position of a position p is treated as the sort order position of a pseudo-argument (p, I). If o is undefined for a value, the function's value is `?`. If s is undefined for a value, the function's value is 0. To keep the notation simple, we assume that the functions' domains are the same when two trees are compared.

For example, Alice's tree in Figure 1e translates to $o = \{(p, \texttt{+}), (a, \texttt{+}), (b, \texttt{+})\}$, $s = \{((a, I), 1), ((a, p), 1), ((b, p), 2)\}$.

A distance function must map the different values to numeric values for calculations. We will evaluate different transformation strategies. As all distance functions need to map the opinion values of O to numeric values and some of them map importance weights to other numeric values, we define the following common mapping strategies:

$$r(x) = \begin{cases} 0.5 & \text{if } x = \texttt{+} \\ 0.25 & \text{if } x = \texttt{(+)} \\ 0 & \text{if } x \in \{\texttt{0}, \emptyset, \texttt{?}\} \\ -0.25 & \text{if } x = \texttt{(-)} \\ -0.5 & \text{if } x = \texttt{-} \end{cases} \tag{1}$$

$$w_h(x) = \frac{1}{x} \tag{2}$$

$$w_g(x) = \frac{1}{2^x} \tag{3}$$

The result of a division by 0 is defined as 0, which means that arguments without importance value (which default to 0) get a calculated weight of 0. The variants $w_{\bar{h}}$ and $w_{\bar{g}}$ are defined the same way, but the values are normalized such that the sum of function values for all arguments with the same conclusion is 1 (or 0, if no argument has a value greater 0). For instance, if we take Alice's tree in Figure 1e again, $w_{\bar{h}}((a, p)) = \frac{\frac{1}{1}}{\frac{1}{1} + \frac{1}{2}} = \frac{2}{3}$. If we mention the function name w, any possible variant can be used (thus, the concrete choice of w is a hyperparameter of the distance function).

Sometimes, we refer to the "simple" opinion, which removes the weight part of the opinion:

$$\text{simple} : O \to \{\texttt{+}, \texttt{-}, \texttt{0}, \texttt{\O}, \texttt{?}\} : x \mapsto \begin{cases} \texttt{+} & \text{if } x = (\texttt{+}) \\ \texttt{-} & \text{if } x = (\texttt{-}) \\ x & \text{otherwise} \end{cases} \tag{4}$$

3 Distance Functions for Argumentations

We now present the distance functions we have compared. Most functions are based on previous work in argumentation theory or related fields and have been adapted by us for use with the formal definition introduced in Section 2. Most functions have hyperparameters, e.g. which function w is used. An overview of the distance functions, their hyperparameters and tested ranges can be found in Table 1.

Table 1: Overview of examined distance functions and their hyperparameters with tested values

Function	Hyperparameters
Bhavsar	$w \in \{w_{\bar{h}}, w_{\bar{g}}\}$, $N \in \{.1, .25, .5, .75, .9\}$
Cosine	$w \in \{w_g, w_{\bar{g}}, w_h, w_{\bar{h}}\}$
Jaccard	$\text{set} \in \{\text{set}_a, \text{set}_o, \text{set}_s, \text{set}_{s'}\}$, $\text{keep} \in \{\text{keep}_a, \text{keep}_t\}$
p-metric	$p \in \{1, 2\}$, $ds \in \{d_{s_w}, d_{s_s}\}$, $da \in \{da_0, da_s\}$, $w \in \{w_g, w_{\bar{g}}, w_h, w_{\bar{h}}\}$
Soergel	$w \in \{w_g, w_{\bar{g}}, w_h, w_{\bar{h}}\}$
VAA	–
WATD	$\alpha \in \{.1, .25, .5, .75, .9\}$, $w \in \{w_g, w_{\bar{g}}, w_h, w_{\bar{h}}\}$

Bhavsar distance [2] presented a metric for match-making of agents in e-business environments, which are represented as trees. As the definition of that recursive metric is lengthy, we do not repeat its definition here. The metric can be applied to our structure by transforming sort orders using $w_{\bar{h}}$ or $w_{\bar{g}}$, and treating opinions as node labels. A parameter N sets the relative importance of subtrees and respective roots, similar to the PageRank algorithm [15].

Cosine distance We define the Cosine distance similar to [16], who predict opinions in argumentation. They treat accepting and declining a statement s as two different entities ("acceptance of s" and "acceptance of $\neg s$") and ignore a statement if it has no rating in one of the inputs:

$$d(t_1, t_2) = 1 - \frac{V_1 \cdot V_2}{||V_1||\,||V_2||} \tag{5}$$

where an argumentation tree $t_i = (o_i, s_i)$ is transformed to a vector V_i with the components $s_i(a)$ for every argument a, and $\max(-r(o_i(s)), 0)$ and $\max(r(o_i(s)), 0)$ for every statement s for which both trees have no `?` opinion.

Jaccard distance The Jaccard distance has been used by [11] as the basis for calculating the similarity of process models. We apply it in the following form:

$$d(t_1, t_2) = \frac{|\mathrm{set}(t_1) \triangle \mathrm{set}(t_2)|}{|\mathrm{set}(t_1) \cup \mathrm{set}(t_2)|} \tag{6}$$

where the functions "set" and "keep" are chosen from

$$\mathrm{set}_a((o,s)) = \mathrm{set}_o((o,s)) \cup \mathrm{set}_s((o,s)) \tag{7}$$
$$\mathrm{set}_o((o,s)) = \{(x,y) \mid (x,y) \in o \land \mathrm{keep}(y)\} \tag{8}$$
$$\mathrm{set}_s((o,s)) = s \tag{9}$$
$$\mathrm{set}_{s'}((o,s)) = \mathrm{simple}(s) \tag{10}$$
$$\mathrm{keep}_a(x) = 1 \tag{11}$$
$$\mathrm{keep}_t(x) = \begin{cases} 1 & x \in \{\texttt{+}, \texttt{(+)}, \texttt{0}, \texttt{(-)}, \texttt{-}\} \\ 0 & \textit{otherwise} \end{cases} \tag{12}$$

If "set_o" is used for "set", argument weights are completely ignored; "set_s" completely ignores opinions and only looks at argument and position weights. "keep" determines if unknown (`?`) and "no opinion"s ($\emptyset$) are included.

The argumentation software Carneades [8] uses a special case of this distance function with set = $\mathrm{set}_{s'}$, which means that the relative number of different opinion tendencies is counted.

***p*-metric** This distance function is based on the p-metric for fuzzy sets [20].

$$d(t_1, t_2) = \left(\sum_{s \in S} ds(o_1(s), o_2(s)) + \sum_{a \in A} da(s_1(a), s_2(a)) \right)^{\frac{1}{p}} \tag{13}$$

with $p \in \mathbb{N}$, and ds, da one of

$$ds_w(o_1, o_2) = \begin{cases} 0 & \text{if } o_1 = o_2 \\ 1 & \text{if } o_1 \text{ or } o_2 \text{ in } \{\emptyset, ?\} \\ |r(o_1) - r(o_2)|^p & \text{otherwise} \end{cases} \tag{14}$$
$$ds_s(o_1, o_2) = |d_w(\mathrm{simple}(o_1), \mathrm{simple}(o_2))|^p \tag{15}$$
$$da_0(s_1, s_2) = 0 \tag{16}$$
$$da_s(s_1, s_2) = |w(s(s_1)) - w(s(s_2))|^p \tag{17}$$

Soergel distance This distance function is also known as weighted Jaccard distance, which has also been used by [16]. We use the following definition, which uses the same vector representation as defined for the Cosine distance above:

$$d(t_1, t_2) = 1 - \frac{\sum_i \min(V_{1_i}, V_{2_i})}{\sum_i \max(V_{1_i}, V_{2_i})} \tag{18}$$

where V_{1_i} is the i-th component of the vector representation of t_1.

VAA distance In many Voting-Advice Applications (VAAs), the distance between a user's attitudes and political party's attitudes on political positions are compared. One possibility is using proximity voting logic [17], optionally weighted, which doubles the influence of a position (as, for example, used by the German Wahl-O-Mat application [13]). We adapted the idea to our model:

$$d(t_1, t_2) = \sum_{p \in P} u(o_1(p), o_2(p)) \cdot v_{t_1,t_2}(s_1(p), s_2(p)) \cdot z(o_1, o_2) \tag{19}$$

with

$$u(o_1, o_2) = \begin{cases} 2 & \text{if } o_1 \text{ or } o_2 \text{ in } \{\texttt{+}, \texttt{-}\} \\ 1 & \text{otherwise} \end{cases} \tag{20}$$

$$v_{t_1,t_2}(s_1, s_2) = \begin{cases} 2 & \text{if } s_1 \text{ or } s_2 \text{ is in the top half (rounded down)} \\ & \quad \text{of the ratings for positions} \\ 1 & \textit{otherwise} \end{cases} \tag{21}$$

$$z(o_1, o_2) = \begin{cases} 0 & \text{if } o_1 \text{ or } o_2 \text{ in } \{\emptyset, \texttt{?}\} \\ |r(\text{simple}(o_1) - r(\text{simple}(o_2))| & \text{otherwise} \end{cases} \tag{22}$$

Note that, as in a VAA, only positions are considered, and statements which are no positions are ignored. Moreover, both arguments and positions can contain weights in our model, whereas a VAA typically only allows voters to input weights.

Weighted argumentation tree distance (WATD) In [3], we have suggested a pseudometric for argumentations with weighted edges and nodes. This metric respects the structure of an argumentation tree by limiting the influence of each branch to its importance, and giving statements deeper in the tree a lower weight. Adapted to the tree model in this paper, the metric is defined for two trees $t_1 = (o_1, s_1)$, $t_2 = (o_2, s_2)$ as follows:

$$d(t_1, t_2) = (1 - \alpha) \sum_{s \in S} \alpha^{\text{de}(s)} \left| \prod_{a \in A_{s \to I}} w(s_1(a)) r(o_1(s)) - \prod_{a \in A_{s \to I}} w(s_2(a)) r(o_2(s)) \right| \tag{23}$$

with $\alpha \in (0, 1)$ (a lower α emphasizes opinion on statements closer to the root, similar to N in the Bhavsar distance), $A_{s \to I}$ the set of all arguments from statement s to the root I, and $\mathrm{de}(s)$ the depth of a statement s, where positions have a depth of 1. This basic idea is to multiply each opinion value of t_1 with the product of all weight from the root node I to that opinion calculate the distance to the same value in t_2. Thereby, opinion difference closer to the root have a higher influence than "deeper" opinions.

4 Comparison with a Human Baseline

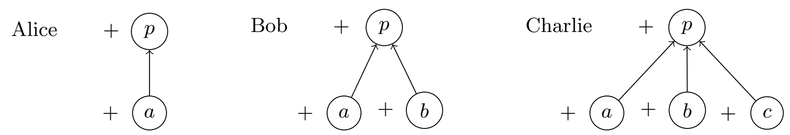
We think that the best way to check whether a distance function is intuitive is comparing it with a human baseline. In an online survey we have previously conducted [5], different possible properties for distance functions comparing attitudes in argumentation settings have been checked for their intuitiveness. In the survey, around 40 assessments by untrained human subjects have been collected for different argumentation scenarios. From the survey results, we can get a list of properties which should be fulfilled by a distance function to match human intuition. If we look only at properties which can be considered intuitive from that survey on a significance level $\alpha = 10\%$[3], we get a list of 17 properties which should be fulfilled.

For many hypotheses, also comparison questions not directly relevant for the hypotheses have been asked in the original questionnaire[4]. For instance, if we wanted to know whether Alice's attitude is more similar to Charlie's or Bob's attitude, we also asked whose attitude is closest to Bob's. For those hypotheses, we also considered properties which can be derived from the additional questions, if they are significant. Those additional properties will be marked with a superscript ^{A}, and all resulting sub-hypotheses are numbered with the according sub-question number (e.g., H2.1^{A} is the first question for the questionnaire scenario for H2).

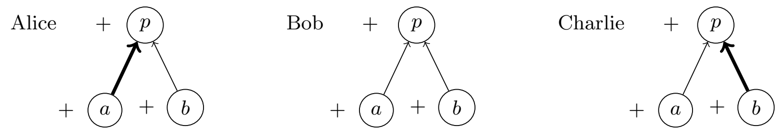
Table 2 lists all relevant hypotheses from [5] which we used as the basis for our comparison. For this paper, we changed the formulation of the hypotheses to match the real outcome of the survey to reflect the actual property expected from a distance function. Note that for H18, two scenarios were used, where only one yielded significant results, which is why this hypothesis has been completely reformulated. Figure 1 depicts visualizations of the concrete questionnaire scenarios for some more complex hypotheses, and also what similarity order is expected, based on the survey results. For example, in Figure 1e, Bob's attitude should have a smaller distance to Alice's attitude than to Charlie's attitude. Since the answers for human intuition are known only for those concrete scenarios, we will only use those concrete examples as the basis for the comparison of distance measures.

[3] including Bonferroni correction for multiple comparisons, i.e. we assure that the type I error rate is less than 10% by requiring p-values less than $\frac{\alpha}{\text{number of possible answers}}$

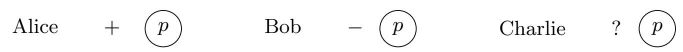
[4] cf. raw data at `https://github.com/hhucn/argumentation-similarity-survey-results/`



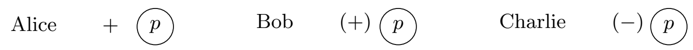
(a) H2.1^{A}: $d(A,B) < d(A,C)$, H2.2: $d(B,C) < d(B,A)$, H2.3^{A}: $d(C,B) < d(C,A)$



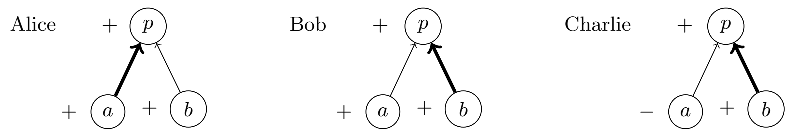
(b) H5: $d(A,B) < d(A,C)$



(c) H8.1^{A}: $d(A,B) = d(A,C)$, H8.2^{A}: $d(B,A) = d(B,C)$, H8.3: $d(C,A) = d(C,B)$



(d) H13.1^{A}: $d(A,B) < d(A,C)$, H13.2: $d(B,A) < d(B,C)$, H13.3^{A}: $d(C,B) < d(C,A)$



(e) : H14.1^{A}: $d(A,B) < d(A,C)$, H14.2: $d(B,A) < d(B,C)$

Alice:	Bob:	Charlie:
1. $b+$	1. $a+$	1. $a+$
2. $a+$	2. $c+$	2. $b+$
3. $c+$	3. $b+$	3. $c+$

(f) H16: $d(C,B) < d(C,A)$

Alice:	Bob:	Charlie:
1. $a+$	1. $a+$	1. $c+$
2. $b+$		2. $b+$
3. $c+$		

(g) H21.1: $d(A,B) < d(A,C)$, H21.2^{A}: $d(B,A) < d(B,C)$

Fig. 1: Visualization of questionnaire scenarios (and thus, test scenarios) of some hypotheses; $d(A,B)$ denotes the distance between Alice and Bob etc.

Table 2: All relevant hypotheses which we included in our comparison. Deviations from the original formulations in [5] are *emphasized*.

#	Property
H2	Proportionally bigger overlap on arguments for/against a position results in greater similarity than the absolute number of differences.
H3	A neutral opinion is between a positive and a negative opinion.
H5	Weights of arguments have an influence even if they are the only difference.
H7	No opinion *has the same distance from* a positive and a negative opinion *if a decision is forced.*
H8	An unknown opinion *has the same distance to a positive and a negative opinion as* a positive and a negative opinion *if a decision is forced.*
H12	It is possible for a difference in arguments for/against positions to result in greater dissimilarity than a difference in opinions on those positions.
H13	Two argumentations with weak and contrary opinions on a statement can *not* be closer than two argumentations with the same opinions, but with very different strength.
H14	Two argumentations with weak arguments and contrary opinions on their premises can *not* be closer than two argumentations with the same opinions, but with very different strength of arguments.
H16	Flipping the two most important positions results in a bigger difference than flipping two less important positions.
H18	*Moving the least important position to the top results in greater dissimilarity than changing the order of item 2 to 4.*
H19	Agreeing with someone's most important position is as important as having that person's most important opinion matching mine.
H20	Adding another most important position *(which is neutral in the other argumentations)* results in greater dissimilarity than flipping the priorities of two positions.
H21	Having more similar priorities of opinions can result in greater similarity even with lower absolute number of same opinions.
H22	Not mentioning a position results in greater dissimilarity than assigning lower priorities.

The following hypotheses have not been considered although our inclusion criterion is fulfilled: A variant of H8, which says an unknown opinion vs. a positive and a negative opinion cannot be assessed, has been excluded, because this would result in a partially defined distance function, which we consider undesirable. H9 only checked text comprehension and has no implications for a distance function. H15, which included an undercut attack, is not included since the original question was probably misleading/not understood by the participants, as discussed in [5].

We now present which distance functions fail on which reference scenarios, and give explanations on why certain distance functions fail on specific cases. We have tested each distance function with every possible combination of hyperparameters with the relevant scenarios. Table 3 summarizes which cases yield the expected results for each distance function with the best parametrization (i.e. maximum number of expected results). Those parametrisations are depicted in Table 4.

The p-metric fails on H21.1 (cf. Figure 1g) only, which happens because the missing weights for b and c have a greater influence than the common most important position of Alice and Bob. The Jaccard distance function also fails on H21.1 since it only considers that Alice and Charlie have more positions in common.

Table 3: Overview of cases fulfilled by the individual distance functions for the parametrisation which yield the highest number of fulfilled cases; ^{e} are failing cases where the calculated distance is 0; the numbers of sub-hypotheses refer to the question number in the original questionnaire.

Hypothesis	Bhavsar	Cosine	Jaccard	p-metric	Soergel	VAA	WATD
H2.1^{A}	✓	✓	✓	✓	✓	^{e}	✓
H2.2	✓	✓	✓	✓	✓	^{e}	✓
H2.3^{A}	✓	✓	✓	✓	✓	^{e}	✓
H3	✓	✓	✓	✓	✓	✓	✓
H7	✓	✓	✓	✓	✓	✓	✓
H8.1^{A}			✓	✓			
H8.2^{A}			✓	✓			
H8.3	✓	✓	✓	✓	✓	✓	✓
H5	^{e}	✓	^{e}	✓	✓	^{e}	✓
H12	✓	✓	✓	✓	✓		✓
H13.1^{A}	✓	✓	✓	✓	✓	✓	✓
H13.2	✓	✓	✓	✓	✓	✓	✓
H13.3^{A}	^{e}	✓	^{e}	✓	✓	✓	✓
H14.1^{A}	✓	✓	✓	✓	✓	^{e}	✓
H14.2	✓	✓	✓	✓	✓	^{e}	^{e}
H16	^{e}	✓	^{e}	✓	✓	^{e}	✓
H18	^{e}	✓	^{e}	✓	✓	^{e}	✓
H19	✓	✓	✓	✓	✓	✓	✓
H20	✓	✓	✓	✓	✓	✓	✓
H21.1	✓	✓			✓	^{e}	✓
H21.2^{A}	✓	✓	✓	✓	✓	^{e}	✓
H22	✓	✓	✓	✓	✓	^{e}	✓
$\sum$	16	20	17	21	20	8	19

Table 4: Best parametrisations for each distance function; each combination of the listed parameters yields the same (best) results.

Function	Best parametrisations
Bhavsar	$w \in \{w_{\bar{h}}, w_{\bar{g}}\}$, $N \in \{.1, .25, .5, .75, .9\}$
Cosine	$w \in \{w_g, w_{\bar{g}}, w_{\bar{h}}\}$
Jaccard	$\text{set} \in \{\text{set}_{s'}\}$, $\text{keep} \in \{\text{keep}_t\}$
p-metric	$p \in \{1, 2\}$, $ds \in \{d_{s_w}\}$, $da \in \{da_s\}$, $w \in \{w_{\bar{g}}, w_{\bar{h}}\}$
Soergel	$w \in \{w_g\}$
VAA	–
WATD	$\alpha \in \{.25, .5, .75, .9\}$, $w \in \{w_{\bar{g}}, w_{\bar{h}}\}$

All cases for H2 (cf. Figure 1a) fail only for the VAA distance function, where equal distances are calculated instead of different ones, because the arguments, which are the only difference in this case, are completely ignored by this function. The same applies to H5, H12, H14.1^{A} and H14.2. H14 also fails for WATD, because the distance function has been designed to *not* fulfill this property [3, Desideratum 7].

Many properties involving changing the importance order of positions, namely H16, H18, H21.1, H21.2^{A}, and H22, fail for the VAA function since it does not have a fine-grained differentiation of importance which is necessary to capture the differences.

All distance measures except for Jaccard and p-metric fail to give a positive and a negative opinion a distance which is equal to the distance to an unknown opinion (H8.1^{A}, H8.2^{A}, cf. Figure 1c). Jaccard is good here because it treats any difference of opinion as equally distant; the p-metric explicitly defines every comparison with an unknown opinion as 1. On the other hand, e.g., WATD is defined to treat an unknown opinion as falling between positive and negative, and the VAA metric ignores a position if the opinion in one graph is unknown.

H5 states that the difference between argumentations should be non-zero even if argument weights are the only difference. This fails for Bhavsar by design of the metric [2, Example 2]. The best parametrisation for Jaccard ignores weights, so it also fails here. For the same reason, H16 and H18 fail for both distance functions.

H13.3^{A} checks that a negative opinion (`-`) is closer to a weak positive opinion (`(+)`) than to a stronger positive opinion (`+`). Bhasvar and Jaccard distance functions fail to see a difference here because they treat the distances between any of the opinions `-`, `(+)`, and `+` the same.

To sum up, Cosine, p-metric, and Soergel yield the best results, matching human intuition in more than 90% percent of the tested cases.

5 Discussion

From our evaluation, one gets an idea which metrics yield intuitive results for applications which compare attitudes in argumentations. Nevertheless, we want to point out some limitations of our comparison method.

Firstly, we did not have a look at bigger argumentation hierarchies, or argumentation with re-used statements (e.g. cycles). For the former, our previous survey did not give significant results, for the latter, no reference data has been collected in the survey because cycles are hard to grasp with intuition. Hence, distance functions which model those cases (e.g. the original WATD pseudometric) have a disadvantage because this feature is not considered in the comparison.

From the survey results, it is also possible to conduct properties which should *not* be fulfilled. There are cases where there is no significant "true" answer, but there are clear "false" answers. Furthermore, the list of properties and cases checked in this paper is probably incomplete and can be extended with additional intuitive properties, which might then change the ranking of distance functions.

As we built upon the results of our previous survey, and we are not aware of similar surveys, we did not include more properties.

Note that we did not check whether the original properties as presented in Table 2 are fulfilled in general, but only whether the concrete questionnaire scenarios yielded the expected, “intuitive” results. We did this because the original survey did not find out whether the hypotheses are true, but only collected results for the specific scenarios. Moreover, all properties get equal weight. Depending on the application (e.g., a VAA), some properties might not be relevant. What is more, some distance functions might get better results if the underlying representation model is changed.

Finally, it will be interesting to evaluate distance functions not on concrete artificial scenarios, but in an application context, e.g. a recommender system, since this might produce different results. A challenge for real applications is retrieving the necessary pieces of information from a user, e.g. how important an argument is considered, within an intuitive user interface.

6 Related Work

There is only limited related research in the evaluation and development of distance function in the context of argumentation, but there are some applications of such distance functions which have been studied.

A dataset with 16 positions on 4 issues has been published by [16]. 309 students gave their opinions on those issues by giving arguments and their level of agreement with that argument on a scale from -1 (total disagreement) to 1 (total agreement). They compare different algorithms for predicting user opinions on positions. A kind of soft cosine measure, where feature similarity is exploited using position correlation, performed best in their comparison. The comparison also included, i.a., collaborative filtering using Jaccard similarity, ordinary Cosine similarity, and other, model-based algorithms, e.g. a neural network.

Their work focuses on the application of measuring similarity in the concrete context of a recommender system, whereas we focus on calculating relative similarities to get a similarity order for user attitudes. Similarly, [18] tested different recommender agents in laboratory argumentation settings. [9] uses collaborative filtering and clustering in a social network context to find political parties closest to a user. The collaborative filtering was used to predict missing values to make clustering with sparse information easier.

Related work in other domains than argumentation chose a similar way of evaluation with a human baseline as we did in this paper.

In the context of word similarity, [12] proposed different distance functions, and compared them with human ratings from a dataset created by [14]. They also indicate that the best way to determine the quality of a distance function is comparing it with human common sense. Within the same application context, [6] agrees that “comparison with human judgments is the ideal way to evaluate a measure of similarity”.

The study presented in [1] is based on the study design of [14]. 50 human subjects assessed the similarity of process descriptions, and compared those assessments with the values of five metrics. The results did not correlate well, but the correlation with the metrics was not worse than the correlation between the human subjects. [11] present a metric based on the Jaccard coefficient for process model similarity. They compared the results of the metric with human assessment in an information retrieval task.

[19] evaluated six different similarity measures (i.a., $l1$, $l2$ norm, pointwise mutual information) with the application in a recommender system for online communities using item-based collaborative filtering. A similarity measure has been considered good if the user wanted to join the suggested community. The $l2$ norm performed best, although the authors found other tested measures, which incorporated mutual information, more intuitive.

7 Conclusion and Future Work

We have presented several distance functions for comparing the attitudes of different persons in an argumentation. We compared the performance of the functions in various scenarios with a human baseline taken from a survey we have previously conducted [5]. The distance functions based on the p-metric, Cosine, and Soergel distance performed best on our dataset. Those results can be used for developing applications which should give results matching human intuition, e.g. when developing a distance-based recommender system for arguments, or clustering of opinions.

For future work, an extended comparison with more scenarios for a human baseline would be useful, i.a. for deeper argumentations. A comparison in different application scenarios can give more insights. We plan to compare different metrics in an argument-based voting advice application in an empirical study. Another aspect for further research is the question of how to gather the information needed from users without having user interfaces which are too crowded.

References

1. Bernstein, A., Kaufmann, E., Bürki, C., Klein, M.: How similar is it? towards personalized similarity measures in ontologies. In: Wirtschaftsinformatik 2005, pp. 1347–1366. Springer (2005)
2. Bhavsar, V.C., Boley, H., Yang, L.: A weighted-tree similarity algorithm for multi-agent systems in e-business environments. Computational Intelligence **20**(4), 584–602 (2004)
3. Brenneis, M., Behrendt, M., Harmeling, S., Mauve, M.: How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs. In: Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020). pp. 2–13. No. 2672 in CEUR Workshop Proceedings, Aachen (Sep 2020)
4. Brenneis, M., Mauve, M.: deliberate – Online Argumentation with Collaborative Filtering. In: Computational Models of Argument. vol. 326, p. 453–454. IOS Press (Sep 2020). https://doi.org/10.3233/FAIA200530

5. Brenneis, M., Mauve, M.: Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations. In: Proceedings of the Workshop on Advances In Argumentation In Artificial Intelligence 2020. pp. 1–15. No. 2777 in CEUR Workshop Proceedings, Aachen (Nov 2020)
6. Budanitsky, A., Hirst, G.: Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and other lexical resources. vol. 2, pp. 2–2 (2001)
7. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (1995)
8. Gordon, T.F.: Structured consultation with argument graphs. From Knowledge Representation to Argumentation in AI. A Festschrift in Honour of Trevor Bench-Capon on the Occasion of his 60th Birthday pp. 115–133 (2013)
9. Gottipati, S., Qiu, M., Yang, L., Zhu, F., Jiang, J.: Predicting user's political party using ideological stances. In: International Conference on Social Informatics. pp. 177–191. Springer (2013)
10. Kunz, W., Rittel, H.W.J.: Issues as elements of information systems, vol. 131. Citeseer (1970)
11. Kunze, M., Weidlich, M., Weske, M.: Behavioral similarity – a proper metric. In: International Conference on Business Process Management. pp. 166–181. Springer (2011)
12. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on knowledge and data engineering **15**(4), 871–882 (2003)
13. Marschall, S.: The online making of citizens: Wahl-O-Mat. The making of citizens in Europe: New perspectives on citizenship education pp. 137–141 (2008)
14. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and cognitive processes **6**(1), 1–28 (1991)
15. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (November 1999), `http://ilpubs.stanford.edu:8090/422/`, previous number = SIDL-WP-1999-0120
16. Rahman, M.M., Sirrianni, J., Liu, X.F., Adams, D.: Predicting opinions across multiple issues in large scale cyber argumentation using collaborative filtering and viewpoint correlation. The Ninth International Conference on Social Media Technologies, Communication, and Informatics pp. 45–51 (2019)
17. Romero Moreno, G., Padilla, J., Chueca, E.: Learning VAA: A new method for matching users to parties in voting advice applications. Journal of Elections, Public Opinion and Parties pp. 1–19 (2020)
18. Rosenfeld, A., Kraus, S.: Providing arguments in discussions on the basis of the prediction of human argumentative behavior. ACM Transactions on Interactive Intelligent Systems (TiiS) **6**(4), 1–33 (2016)
19. Spertus, E., Sahami, M., Buyukkokten, O.: Evaluating similarity measures: A large-scale study in the orkut social network. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 678–684 (2005)
20. Xuecheng, L.: Entropy, distance measure and similarity measure of fuzzy sets and their relations. Fuzzy sets and systems **52**(3), 305–318 (1992)

Table 3.1: Overview of cases fulfilled by the individual distance functions for the parametrisations which yield the highest number of fulfilled cases (same as in Subsection 3.4.2) for statistically less significant cases; the carry refers to the sum for statistically more significant cases presented in Subsection 3.4.2.

Hypothesis (p-value)	Bhavsar	Cosine	Jaccard	p-metric	Soergel	VAA	WATD
H1.1$^{(A)}$(.073)	✓	✓	✓	✓	✓	^{e}	✓
H1.2 (.073)	✓	✓	✓	✓	✓	^{e}	✓
H1.3$^{(A)}$(.020)	✓	✓	✓	✓	✓	^{e}	✓
H10.1$^{(A)}$(.059)	✓	✓	✓	✓	✓	^{e}	✓
H10.2 (.059)	✓		^{e}	^{e}		^{e}	✓
H10.3$^{(A)}$(.004)	✓	✓	✓	✓	✓	^{e}	✓
H11 (.174)						✓	✓
carry	16	20	17	21	20	8	19
$\sum$	22	25	22	26	25	9	26

3.4.3 Additional Findings and Implications For Our Pseudometric

In our comparison, we only included hypotheses which had significant results at a level of $\alpha = 10\%$, which meant we considered only answers which were significant with $p < \frac{10\%}{3}$ for most questions, as most of them had 3 answer options. We now examine how the results change if we also look at some statistically less significant cases with $p < 20\%$.

When we take into account those additional hypotheses with higher p-values and their associated additional properties, we see that our pseudometric and the simpler p-metric become the best distance functions, as depicted in Table 3.1. Our pseudometric is now in a par with the p-metric since the former fulfills all the new cases.

The p-metric fails in two of those cases: H10.2 (*Argumentation differences in a branch with lower importance contribute less to dissimilarity* when being against an argument.) fails because the argumentation hierarchy is not considered. Hence, the branch importance cannot have an influence on deeper statements in this distance function. H11 (*Disagreeing on a position results in greater distance than having the same opinion on that position, but with contrary arguments.*) fails due to a similar reason: All opinions have the same weight, regardless of their depth in the argumentation tree. Our pseudometric, on the other hand, uses the factor α to give opinions with smaller depth a higher weight.

Those extended considerations stress that the choice of the correct distance functions also depends on the application context. If there are no deeper argumentation hierarchies in the application, e.g. because only top-level arguments are collected for a single position, using the simpler p-metric might be more suitable than our pseudometric. On the other hand, if we consider, for instance, a VAA where different positions (which may be of different relevance) and their arguments are considered, our pseudometric might be the better choice. In particular, we consider situations like Hypothesis 11, where people agree on the conclusion, but have different arguments, typical of a VAA setting. Also, VAAs should intuitively consider an opinion on a single position more relevant than a single argument for/against a position.

Another aspect is the "turnover-point" for the value of α in our pseudometric, which we already started discussing in Section 3.3. Hypotheses 11 and 12 give extreme examples and we know the turnover-point must be somewhere in between. If we consider a parametrisation with normalization, α must be grater than 0.2 to fulfill Hypothesis 12 (i.e. to allow that many different arguments weigh more than same opinions on positions). To conform with Hypothesis 11 (i.e. to give a single position opinion a higher weight than its arguments), α has to be smaller than 1. This is a great range which should be narrowed down with further experiments.

Chapter 4

Applications for Metrics Comparing Argumentations

We are now well-informed about how attitudes of different participants in an argumentation can be compared. In this chapter, we explore where distance functions which perform that kind of comparison can be applied in the real world.

First, we have a look at our argumentation application *deliberate* (Brenneis and Mauve, 2020a), which uses neighborhood-based collaborative filtering for pre-filtering arguments. The goal is presenting a clearer and less crowded list of arguments to a user, instead of presenting all available arguments from a discussion at once. We also expound on an empirical study as part of the Manchot Research Group Artificial Intelligence project *AI support for policy decisions* (UPEKI), where different recommender algorithms of *deliberate* and their influence on the formation of opinions were investigated. Afterwards, we address how the data collected in the UPEKI project can be used to evaluate recommender systems for arguments and see that our method is better than a simple baseline algorithm.

Last of all, we introduce our argument-based *Voting Advice Application* (VAA) ArgVote (Brenneis and Mauve, 2021b), which calculates voter–party similarity based on their stances on political positions and arguments for/against them. Within an experiment, we could not find out that our way of considering arguments in the matching algorithm is better than not including them, but displaying arguments significantly improved the understanding of political issues and different opinions. In addition, we look at the concept for a VAA chat bot.

4.1 Recommending Arguments With Collaborative Filtering

Many discussion applications like kialo or D-BAS can get confusing for the user when there are too many arguments and the user wants to find new and interesting content. Reading and considering all of them is not feasible and maybe also boring if known arguments or arguments the user considers untrue are read. It would be useful if the arguments presented to the user were pre-filtered in a sensible way. Thus, we have developed the web application *deliberate*, which displays only a subset of available arguments to a user, which contains the arguments the user is likely to accept. Those arguments are obtained using a neighborhood-based collaborative filtering algorithm, based on our pseudometric defined in Section 3.2.

We will now expound on what *deliberate* is, how it works, and how we used it. After a general introduction to the application, its filtering algorithm is explained. We subsequently look at how *deliberate* was used within the UPEKI project and communicate what we learned from the practical application within the empirical studies of the project. Finally, we present a dataset obtained from the UPEKI project which can be used to evaluate argument recommender systems. Using this dataset, we can actually measure that *deliberate*'s recommendation algorithm is better than a simple baseline method.

4.1.1 Introduction of deliberate

For a first overview, we look at the paper for a demo session at which we introduced *deliberate*:

Markus Brenneis and Martin Mauve.
"deliberate – Online Argumentation with Collaborative Filtering"
In: *Proceedings of the 8th International Conference on Computational Models of Argument*, Volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 453–454, IOS Press.
Acceptance Rate: 100%

We made the following contributions in this peer-reviewed paper:

1. introduction of the general idea of *deliberate*
2. rough description of the argument recommender algorithm

The final goal of *deliberate* is reducing the amount of arguments a user has to read by pre-filtering them. The application's back end relies on D-BAS as an argumentation database. This design allows sharing the same argumentation graph with different applications which use D-BAS' *application programming interface* (API).

Personal Contribution

Markus Brenneis, the author of this thesis, wrote the software presented in this work. He made the first drafts and discussed the functionality needed for the UPEKI project with Maike Behrendt, Katharina Gerl, Ole Kelm, Florian Meißner, and Gerhard Vowe. The visual design was created by an external designer and implemented by Markus Brenneis. The paper was written by Markus Brenneis; Martin Mauve provided feedback on the presentation of the key ideas.

Importance and Impact on This Thesis

This publication describes the first practical application of our pseudometric introduced in Section 3.2. The pseudometric is used to calculate the distance of user profiles as part of a neighborhood-based argument recommender system. With *deliberate*, we can check whether our pseudometric is not only interesting from a theoretic point of view, but also viable in practice. *deliberate* plays a key role in the UPEKI project, which we expound on in Subsection 4.1.3.

The final publication is available at IOS Press through http://dx.doi.org/10.3233/FAIA200530.

deliberate – Online Argumentation with Collaborative Filtering

Markus BRENNEIS [a,1], Martin MAUVE [a]

[a] *Department of Computer Science, University of Düsseldorf, Germany*

Abstract. We demonstrate *deliberate*, a full-stack web application to exchange arguments with other users. Collaborative filtering utilizing a specialized metric, which considers the structure of the argumentation tree, is used to suggest arguments which the user is likely to accept.

Keywords. online argumentation, artificial intelligence, collaborative filtering

1. Introduction

Exchanging arguments and keeping track of counter-arguments is important in a world of filter bubbles. *deliberate* is a tool which focuses on providing a broad overview of arguments to reduce the bias due to selective exposure, reduce insecurity about one's opinion, and possibly also change one's opinion when seeing other arguments.

A new concept in our application is pre-filtering the presented arguments using algorithms which use collaborative filtering to show arguments the user will probably accept.

2. *deliberate* – A (Neutral?) Webapp for Exchanging Arguments

deliberate is built around a central statement which is being discussed. The user is first asked for their initial opinion on it, how sure they are about their opinion, and what their most important argument is. They can select an argument from a list of arguments already given by other users, search the database of all arguments, or add a new one, which is similar to other applications for online argumentation.

Using the collected information about the user's opinion, more pro and/or contra arguments previously provided by other users are suggested. The user can indicate that they like or dislike these arguments, sort their arguments by importance, and go deeper into the argumentation graph by selecting a statement. The argumentation graph is based on the IBIS model [2], where nodes are statements and edges are arguments, but the user has not to be aware of this theoretical background.

Unlike similar applications, every list of suggested arguments is pre-filtered using collaborative filtering, which has several advantages. The user only sees arguments which

[1]Corresponding Author: Markus Brenneis, Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany; E-mail: markus.brenneis@uni-duesseldorf.de; member of the Manchot research group *Decision-making with the help of Artificial Intelligence*, use case politics.

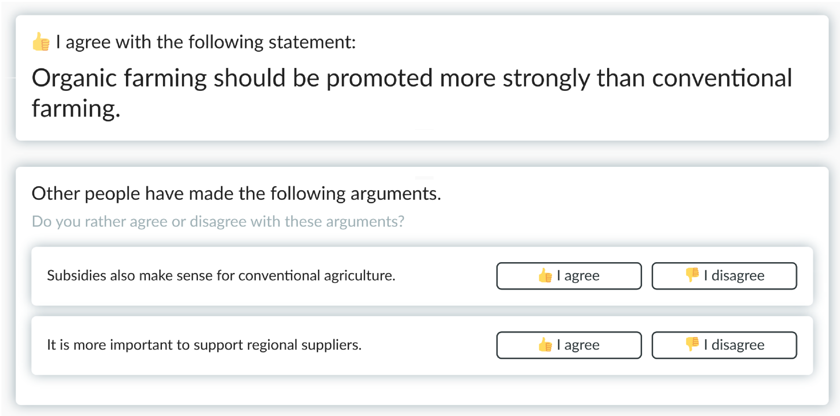


Figure 1. Screenshot of *deliberate*, depicting confrontation with other arguments in a confrontational mode.

are more relevant for them first; thus, they have to read less text, can concentrate on personally relevant arguments, and do not have to read arguments which might be uninteresting.

The filtering uses a new pseudo-metric which takes the characteristics of an argumentation graph into account. For instance, it considers opinions for arguments deeper in the graph as less important, takes into account which arguments are used, and which arguments are rated more important for one's opinion than others. Using this metric, users which are most similar to oneself are determined, and a weighted-average of those users' opinions is calculated. The arguments which have the highest agreement in this average are displayed first.

In a currently running study, we are evaluating the effects of different filtering methods (including and excluding collaborative filtering, showing only arguments against one's own opinion, and others) on the formation of opinion and perception of neutrality.

3. Related Work

In kialo[2], users can exchange arguments in hierarchical pro/con lists, where arguments are sorted by impact, but unlike in our application, the lists are not pre-filtered or sorted based on the users' profile. The mobile application introduced in [1] uses collaborative filtering to predict the agreement of a user with a not yet rated statement; they use, however, a simpler cosine metric which does not incorporate the graph structure, and do not use it for pre-selecting the arguments displayed.

References

[1] Althuniyan, N., Sirrianni, J.W., Rahman, M.M., Liu, X.F.: Design of mobile service of intelligent large-scale cyber argumentation for analysis and prediction of collective opinions. In: International Conference on AI and Mobile Services. pp. 135–149. Springer (2019)

[2] Kunz, W., Rittel, H.W.J.: Issues as elements of information systems, vol. 131. Citeseer (1970)

[2] `https://www.kialo.com/`

4.1.2 Description of the Filtering Algorithm

After having got a general idea of how *deliberate* works, we now explain how the argument recommender system works and which related implementations exist. The pre-filtering process has been designed to display arguments which the user most likely agrees to. To achieve this goal, a neighborhood-based collaborative filtering (Elahi et al., 2016) algorithm is used.

The nearest-neighbor algorithm predicts the agreement rating r for an argument's premise p. In general, the algorithm works like this:

1. Take all other users for whom the rating of p is known.
2. Sort the users according to the distance d to the current user calculated by our pseudometric (cf. Section 3.2).
3. Keep the n nearest users.
4. Calculate the average rating for the premise by the remaining users, weighted by $M - d$, where M is the maximal possible distance (cf. Subsection 3.2.1).

Within this calculation, there are two hyperparameters: the number n of nearest users to consider and the parameter α of the pseudometric. We will evaluate which choices of the parameters are good in Subsection 4.1.5.

Depending on how *deliberate* is configured, arguments with a high or low predicted rating, randomly chosen arguments, or arguments against one's own opinion can be displayed. We explain the idea behind those different modes in Subsection 4.1.3.

A similar, but different objective was tackled before by Rahman et al. (2019). They collected a dataset with 4 issues and 16 positions. Users stated their attitudes by adding arguments and they indicated how much they agreed with each argument. Using that information, different algorithms predicting the agreement with the position were evaluated. Thus, the opposite task was considered here, since we are interested in predicting the opinion on arguments, not the opinion on positions.

Opponent modeling (Hadjinikolis et al., 2013) is a related, but broader field. An opponent model does not only comprise information on an agent's beliefs and preferences, but also abilities, objectives, and strategy. A recommender algorithm as presented above might have applications in opponent modeling, since it could predict argumentation behavior.

4.1.3 Application of deliberate Within the UPEKI Project

Our application *deliberate* played a major role in the UPEKI project. The general goal of this interdisciplinary research project is to find out how different selection algorithms for arguments presented to a user influence their formation of political opinions. We now briefly summarize the UPEKI experiments and how *deliberate* was used. For more detailed information about the studies of the UPEKI group and its findings, see Neumann et al. (2021) and Kelm et al. (2021)[1].

The main experiment comprised three pretests (P1, P2, P3) and the main study, which was a panel with three waves (W1, W2, W3; i.e. the same set of people participated three times). *deliberate* was used in two pretests, W2, and W3. Two different positions were discussed: plastic packaging in W2 and genetic engineering in W3. In W2, 3115 participants used *deliberate*; during the first two pretests, there were 167 users. The users were a representative, random sample of the German online population.

Another, smaller experiment conducted between P3 and W1 dealt with the anti-corona measures in Germany. In this study, 276 participants used *deliberate*.

The original design of *deliberate* allowed a free navigation within the argumentation and free interactions with arguments, i.a. one can add arguments and state one's own opinion at any time. But within the UPEKI project, a more "guided" version of *deliberate* was used, which presents those steps one after another. This guidance assures that participants provide all pieces of information needed for the experiment and that there is some kind of "end," where the users can return to the main questionnaire. What is more, the sorting of arguments by importance was replaced by a Likert scale.

The pretests and W1 were used to solve the cold-start problem (Schafer et al., 2007) of recommender systems. Within the pretests, participants provided their opinions on randomly selected arguments and the positions; those pretest user profiles could then be used by the recommender algorithm in W2 and W3. In W1, participants were asked to provide their opinions on and ratings for the two positions later discussed in W2 and W3 to initialize their user profiles.

Within the project's main study, we studied different filtering algorithms, which were varied in two dimensions:

- direction of arguments: 6 arguments supporting the position, 6 arguments against the position, or 3 supporting + 3 attacking arguments
- argument selection: randomly selected arguments, randomly selected argument (read-only), ai selected arguments

Here, "read-only" means that no further interaction with the arguments presented content is possible, i.e. the user can neither indicate their opinion on the argument presented, nor add new arguments.

We are aware that any kind of pre-selection of arguments can be used to manipulate users. We think that whatever algorithm is used in a real application (i.e. outside an experiment which studies the influence of different algorithms) should be transparent to the user and, in the best case, configurable by the user. Furthermore, users should always have the option to see all available arguments to get an unbiased view.

[1] Note that *deliberate* is called *Discuss!* within the UPEKI project. In this thesis and all publications included, the name *deliberate* is used to prevent confusion with *discuss* by Meter et al. (2017).

Table 4.1: Comparison of the user-friendliness of *deliberate* ($n \in \{2662, 2938\}$), D-BAS (average $n = 23$), and *discuss* ($n = 35$) on a Likert scale from "absolutely disagree" (1) to "absolutely agree" (5).

Aspect	deliberate	D-BAS	discuss
enjoyed application	3.48	**3.81**	
understood application	**4.01**	3.80	3.65
no problems with navigation	**3.77**	3.38	3.60
got lost	**1.89**		2.50
would use application again	3.48	4.35	**4.40**

4.1.4 Findings From the Field

Within the questionnaires of the UPEKI project, several questions on the usability of *deliberate* and the influence on the subjective and objective level of information were asked. We now summarize the key findings regarding *deliberate* we could get from this data and what we can learn for the design of other argument-based software, e.g. our VAA, which is presented in Section 4.2.

Our questions on the user-friendliness of *deliberate* were based on similar questions asked in the evaluation of the argumentation systems D-BAS and *discuss*. Note that the formulation of questions was not exactly the same, partly because the UIs of the applications is different, and the groups of participants are different. Nevertheless, a rough comparison is possible.

In comparison to D-BAS, users enjoyed *deliberate* a bit less. This might be due to less freedom and more confrontation with other opinions within *deliberate*.

deliberate was, on average, easier to understand than D-BAS and *discuss*; a possible reason for this is that a very "guided" version of *deliberate* was used, where the users were given clear instructions on what to do. Probably for the same reason, there were fewer problem with the navigation in *deliberate* and fewer people got lost. We can learn from this that dealing with arguments and navigating through an argumentation tree is not too hard for users.

On the other hand, fewer people would use *deliberate* again for a similar topic compared to D-BAS and *discuss*. A possible explanation is that the participants in the studies with D-BAS and *discuss* were motivated to argue and they identified with the topics used, since the argumentation aspect and the topic were communicated before study participation. Furthermore, those other studies targeted university students, who might be more interested in discussing a topic in general, whereas the study involving *deliberate* had a representative online sample. We noted, however, that the participants with university degree were least willing to use *deliberate* again.

From UPEKI pretests, we also learned that people can have problems with identifying whether an argument supports or attacks a position. An evaluation of pretest data by Ole Kelm revealed, for example, that people who are against genetic engineering only recognized 4 of 20

Table 4.2: Perceived hardness of the tasks in *deliberate* on a Likert scale from "too hard" (1) to "too easy" (5), $n \in [122, 159]$.

Task	Assessment
give reason for my opinion on arguments	2.26
mark argument as (not) convincing	1.99
indicate strength of opinion on position	1.98
indicate strength of argument	1.87
give arguments for/against position	1.81
indicate opinion on position	1.68

arguments supporting genetic engineering as *supporting* arguments. Conversely, people in favor of genetic engineering recognized all supporting arguments, but only 8 of 20 counterarguments. He found similar correlations for the position on prohibiting plastic packaging, but no such correlation in arguments for and against anti-corona measures. We think that one explanation for the latter is that opinions regarding the coronavirus are not yet "final" for many people, so they feel less uncomfortable when exposed to arguments against their own view.

Based on those findings, we conclude that a transparent application should clearly indicate whether an argument is supportive or attacking within the UI. If that level of transparency is not desirable because of an experiment's goal, the formulations of arguments must be made very clear, as it has been done in the UPEKI main study.

Pretest participants were asked to rate how hard they considered the tasks they had to do within *deliberate*. From the results depicted in Table 4.2, we can deduce that giving one's own arguments for/against other arguments was considered the simplest of all tasks, whereas indicting one's own opinion on the main position was regarded hardest. It is interesting to see that other tasks like indicating how strong an opinion or an argument is felt simpler for many participants. When considering that a VAA's main task is to get the opinion on positions, any additional tasks seem not to be harder for users. Note, though, that increasing the number of tasks makes using the overall application more uncomfortable, thus having a conscious and clear UI is important.

One goal of applications like *deliberate* is increasing what people know about an issue. Hence, we asked people different questions to get to know how informed they feel about the issue presented to them. From the results in Table 4.3a, we see a slightly higher level of information for the participants who used *deliberate*, which is the desired tendency.

After the use of *deliberate*, participants also had to provide different arguments for and against a position argued about before as free text. Here, our results in Table 4.3b indicate that users of *deliberate*, on average, provided fewer arguments than the control group, which is surprising. This might happen because *deliberate* already asked users to enter arguments, thus users were less willing to provide them again. Also, note that the number of arguments has only been counted by one person, the author of this thesis, to get a general tendency for this figure.

Table 4.3: Subjective and objective level of information of users who used *deliberate* and who did not use it.

(a) Subjective level of information on a Likert scale from "absolutely disagree" (1) to "absolutely agree" (5).

Question	deliberate ($n = 2580$)	Control Group ($n = 252$)
could recall facts	**3.69**	3.63
understood complexity of topic	**3.86**	3.80
understood main points	**3.92**	3.79
could discuss the topic	**3.80**	3.71
feel informed well	**3.70**	3.67

(b) Average number of arguments provided for a position in a free text answer in a pretest.

Arguments for	deliberate ($n = 85$)	Control Group ($n = 32$)
support of organic farming	0.82	**0.97**
support of conventional farming	1.07	**1.13**

Finally, we briefly consider whether the use of *deliberate* changes a person's opinion. For plastic packaging, 16% of the users changed their opinion from W1 to W2 (where *deliberate* has been used for this topic), and 15% from W2 to W3. For genetic engineering, 13% changed their opinion from W1 to W2 (no use of *deliberate* for this topic), and 15% from W2 to W3 (after using *deliberate*). From these figures, we cannot deduce a significant effect of the exposure to arguments within *deliberate*.

Kelm et al. (2021) will provide further details and findings from the UPEKI experiment.

4.1.5 Evaluating the Performance of Argument Recommender Systems

In Subsection 4.1.3, we have presented the UPEKI project which uses *deliberate*'s recommender engine in different modes to present arguments to a user. We made a best-guess effort to choose a sensible algorithm because we were not aware of any datasets we could use to evaluate our algorithm on. But the UPEKI main experiment provided us with argumentation data from thousands of participants, which could be used to evaluate different kinds of recommender algorithms.

In this subsection, we deal with our peer-reviewed publication by Brenneis et al. (2021), in which we presented this dataset:

Markus Brenneis, Maike Behrendt and Stefan Harmeling.

"How Will I Argue? A Dataset for Evaluating Recommender Systems for Argumentations"

In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 360–367, Association for Computational Linguistics.

Acceptance Rate: ∼40%

In this paper, we made the following contributions:

1. presentation of a dataset with more than 900 arguments and personal attitudes of more than 600 subjects
2. definition of three different recommender tasks
3. baseline results from a majority baseline algorithms and *deliberate*'s algorithm

As a result, we saw that *deliberate*'s algorithm was better than a simple baseline, but it can still be improved.

Personal Contribution

Markus Brenneis, the author of this thesis, had the original idea for releasing the dataset together with proposals for different challenges. He compiled the final dataset, which is based on data jointly collected with the researchers of the UPEKI project. Marc Feger provided the English translation. Markus Brenneis formulated the tasks and discussed the ideas with Maike Behrendt and Stefan Harmeling. The benchmark was programmed and evaluated by Markus Brenneis. The introduction of the paper was written by Maike Behrendt; she and Markus Brenneis jointly wrote the description of the dataset. The rest of the paper was written by Markus Brenneis. Stefan Harmeling provided feedback on drafts of the paper.

Importance and Impact on This Thesis

The dataset presented in this paper is the result of the first big application of our software *deliberate* presented in Subsection 4.1.1. The data can be used to evaluate the performance of different argument recommender systems. We do this evaluation, i.a., with the nearest-neighbor algorithm explained in Subsection 4.1.2, which is based on our pseudometric from Section 3.2. As our dataset is the first of its kind, it enabled the first quantitative evaluation of our recommender algorithm and showed that it was better than a plain baseline algorithm.

How Will I Argue? A Dataset for Evaluating Recommender Systems for Argumentations

Markus Brenneis
Heinrich-Heine-Universität
Markus.Brenneis@hhu.de

Maike Behrendt
Heinrich-Heine-Universität
Maike.Behrendt@hhu.de

Stefan Harmeling
Heinrich-Heine-Universität
Stefan.Harmeling@hhu.de

Abstract

Exchanging arguments is an important part in communication, but we are often flooded with lots of arguments for different positions or are captured in filter bubbles. Tools which can present strong arguments relevant to oneself could help to reduce those problems. To be able to evaluate algorithms which can predict how convincing an argument is, we have collected a dataset with more than 900 arguments and personal attitudes of 600 individuals, which we present in this paper. Based on this data, we suggest three recommender tasks, for which we provide two baseline results from a simple majority classifier and a more complex nearest-neighbor algorithm. Our results suggest that better algorithms can still be developed, and we invite the community to improve on our results.

1 Introduction

Argumentation is an important tool of human communication and interaction. Arguments allow us to justify our views and opinions and persuade others. They also play an important role when it comes to decision-making. Not only in terms of law and justice (Collenette et al., 2020; Bench-Capon and Modgil, 2009), but also for each and every personal decision we make on a daily basis.

Taking a position on a controversial issue can be difficult, especially when there are many pro and contra arguments to consider. Finding the arguments that are most important and convincing for oneself is an important aspect in the process of decision-making. For a wide range of fields, recommender systems already facilitate our decisions, using collaborative and content-based filtering algorithms (Schafer et al., 2007), filtering the great load of information that can be found online (Bobadilla et al., 2013). A recommender system for argumentations could help users to make decisions more confidently and also gain a better understanding of the whole issue discussed. First applications like the *Predictive and Relevance based Heuristic agent* (Rosenfeld and Kraus, 2016) and our platform *deliberate* (Brenneis and Mauve, 2020) were presented to address this task. They try to present arguments to users which are most relevant for them.

But large-scale datasets to systematically test and evaluate such recommender systems for argumentations outside a laboratory setting are missing. In this work, we provide a dataset including more than 900 arguments and 600 user profiles, obtained as part of a larger study on political opinion-forming. In this study, we let participants interact with our platform *deliberate*, exposing them to arguments we gathered beforehand, concerning two different controversial questions on nutrition policy. The participants could rate the overall strength of the displayed arguments, indicate whether they find them convincing, and add own arguments. They were exposed to the topics at different points of time, such that the user profiles grow over time and the dataset can be used to test predicting future user behavior.

The dataset we provide here should serve to test and evaluate metrics and algorithms for argument recommender systems. As a baseline, we provide our results from two different algorithms on three different tasks which are predicting the user conviction towards an argument, the assigned strength of an argument, and the top-3 convincing arguments. The baseline results are obtained using a plain majority classifier and the existing recommender algorithm of *deliberate* to test its performance. To our knowledge, we provide the first large-scale dataset on the task of argument recommendation which contains user attitudes at different points of time.

The paper is structured as follows. In Section 2, the theoretical basics on argumentation and the terms used in this paper are defined. The data we collected is described in detail in Section 3. Section 4 introduces the three challenges and subtasks for argument recommendation we propose in this work, for which we provide two baseline results which are subsequently discussed. Section 5 gives an overview of related research, and finally, we summarize our work and look at future work.

2 Definitions

In this paper, we use terms based on the IBIS model (Kunz and Rittel, 1970), but our dataset can also be interpreted in bipolar Dung-style (Dung, 1995) argumentation frameworks. The atomic building blocks of argumentations are textual *statements*. Two statements, called *premise* and *conclusion*, form an *argument*. The premise can either support or attack the conclusion. A controversial statement which is argued about is called *position*, e.g., "plastic packaging for fresh food should be prohibited," and is typically an action which can be performed. Positions do not have a conclusion, but they can be used as conclusions when arguing why the position is sensible or not.

All statements define an argumentation graph where statements are nodes and the edges are arguments, i.e., they represent the argumentative relation between statements. For simplicity, user-interfaces like *deliberate* often call the premises themselves *arguments* to hide the technical definition of *argument* from the user. When the conclusion talked about is fixed, an argument can be uniquely identified by its premise.

Individual persons can have different *opinions* on the statements in an argumentation, e.g., agree or disagree with them with different *strengths* (i.e., the person can be (un)sure about their opinion). In real-world applications, a person's opinion on a statement can be unknown, leading to sparse data.

Furthermore, a person can consider an argument more or less convincing than another argument with the same conclusion; we call this *weight*, and we use a value from the interval $[0, 6]$ to represent it, where higher values correspond to stronger weights; this interval directly corresponds to the Likert scale we used during data collection.

We call the collection of weights and opinions of a person in an argumentation *attitude*. A person's attitude and their user name form a *user profile*.

S will refer to a set of statements. For a statement $s \in S$ which is an argument's premise, $c(s) \in \{0, 1\}$ indicates whether the argument is considered convincing (1) or not (0) by a user, and $w(s) \in [0, 6]$ is the associated weight. Predicted values for conviction and weight produced by a prediction algorithm are referred to as $\hat{c}(s)$ and $\hat{w}(s)$, respectively. The set of all user profiles is called U and can be represented as big sparse matrix with user profiles in the rows (i.e., in our case, with columns for the user name, position agreement strength, and, for each argument, columns for premise conviction and argument weight). Table 1 summarizes our notation.

3 Description of the Dataset

We present our new argumentation dataset with arguments on two different positions on nutrition policies in Germany (see Table 2): The prohibition of plastic packaging and the prohibition of genetic engineering. In contrast to other argumentation corpora, we also include the opinions and argument weights of different persons gathered at different points of time as part of an empirical study on political opinion-forming using our argumentation tool *deliberate* (Brenneis and Mauve, 2020).

The two discussed issues have been identified as the most topical and polarizing ones from a preselected set of controversial questions through a pre-test survey before our main study. In the original main study, we examined whether the use of artificial intelligence methods to pre-select arguments participants can see has an impact on the political opinion forming of individuals in the field of nutrition policies.

Now, we first explain the general data collection and the demographics of the participants. Afterwards, we expound on the pieces of information collected for our data set. Finally, we explain how the dataset looks like and where to obtain it.

3.1 Data collection & Participants

The main study was carried out over a period of four months, including three waves of data collection in August 2020 (T_1), October 2020 (T_2) and December 2020 (T_3). A pretest was conducted in April 2020 (T_0). The study participants were selected from the German online population, representative regarding age, gender, and education, and have agreed to the data publication. For the recruiting process and conducting our online study, we commissioned a German market-research company.

Table 1: Notation used throughout the paper.

S	set of statements
$c(s)$	individual's conviction in argument given by premise s (0 or 1)
$w(s)$	individual's integer conviction weight for corresponding argument (0–6)
$\hat{c}(s)$	algorithm's prediction for $c(s)$
$\hat{w}(s)$	algorithm's prediction for $w(s)$
U	set of user profiles
S_u	subset of statements for which the ratings of user u are known
$T_1 \rightarrow T_2$	predicting data from T_2 using data known at time point T_1
$T_2 \rightarrow T_3$	predicting data from T_3 using data known at time point T_2

In total, we had 674 participants whose data is included in our dataset: 264 in the pre-test T_0 and 410 in T_1, from which 121 dropped out in T_2 and 60 in T_3. The age span reaches from 18 to 74 with an average age of 46.5, which is slightly above the average age (44.5 (Statistisches Bundesamt (Destatis))) of the German population. 52.23% of the study participants were male (in comparison to 49.35% in the German population (Statistisches Bundesamt (Destatis))), 47.48% female (50.65% in the population). 42.14% had at least a high school degree, which exceeds the average for the population as a whole where only 33.5% have at least a high school degree (Statistisches Bundesamt (Destatis)).

Besides working with the argumentation tool, participants were presented a questionnaire which embedded the discussion software and collected, i.a., demographic information.

3.2 Data Collected by Us

Throughout each wave, the participants were exposed to arguments concerning the two different issues on nutrition policies. For each position discussed, a set of at least 18 supporting and 18 attacking arguments has been provided by us beforehand. We chose the arguments from a pre-selection of arguments on both topics that were clearly identifiable as pro or con in a pre-test. Other arguments could be added by the participants and the participants provided their attitudes on these positions and arguments.

For example, one statement arguing in favor of genetic engineering which was provided by us is "Genetic engineering is used to improve plants just like classical breeding, which is not prohibited." Participants who were presented that statement as supporting argument had to indicate *whether* they consider this statement to be a convincing argument for genetic engineering (binary decision) and *how much* they are convinced (Likert scale from *not convincing at all* (0) to *very convincing* (6)).

Overall, the following pieces of information were collected:

- T_0: Pre-test data with 264 participants; opinions and opinion strengths on positions about *plastic packaging* and *genetic engineering*; attitudes on at least 7 randomly selected arguments per topic.
- T_1: first main experiment with 410 participants; attitudes (opinions and opinion strengths) on *plastic packaging* and *genetic engineering* (no arguments involved).
- T_2: second main experiment with 289 participants (subset of users from T_1); attitudes (i.e. opinions and weights) on *plastic packaging* and on 3 randomly selected supporting, and 3 randomly[1] selected attacking arguments; users were able to contribute own arguments for/against the issue or other arguments (which were not included in the randomly selected arguments); attitude on *genetic engineering* (possibly changed since T_1).
- T_3: third main experiment with 229 participants (subset of users from T_2); attitudes on *genetic engineering* and 3 randomly selected supporting and 3 randomly selected attacking arguments; users were again able to contribute own arguments; attitude on *plastic packaging*.

To clarify, the settings in T_2 and T_3 only differ in the position being argued about. The opinions on all positions (whether a participant is for allowance or prohibition and how strong their opinion is) have

[1] Due to a technical problem, 8 of 36 arguments were not included in the random selection.

Table 2: Positions and number of records in the dataset; the number of arguments is split in the number of arguments provided by us beforehand and the number of new arguments entered by users (each counted as the number of unique premise statements).

Position	Number of Arguments	No. of User Profiles T_0	T_1	T_2	T_3
Should plastic packaging for fresh food such as fruit and vegetables be allowed or prohibited in Germany?	36+521	264	410	289	
Should the growing of genetically modified plants for food production be allowed or prohibited in Germany?	38+351	264	410		229

been collected at every time point, i.e. it was possible for participants to change their minds between each poll.

Arguments added by the users could be directly for/against the position discussed, or for/against other arguments.

Having collected the data at different points of time has several practical advantages: First, the data from T_0 and T_1 can be used to tackle the cold-start problem (Schafer et al., 2007) when predicting attitudes from T_2 and T_3, since the users' opinions on the positions is known from T_1. What is more, we can realistically check the performance of a real-world recommender system over time: The dataset considers that we might have incomplete information about persons (e.g., no argument attitude information for the new users in T_1), and we take into account that people might change some of their attitudes over time.

3.3 Content of the Dataset

Our complete dataset is freely available online[2] as CSV files, and the argumentation data is also provided in AIF (Chesnevar et al., 2006) for easy use in standard applications for argumentation frameworks. The dataset published in this work is part of a larger dataset with more experimental groups; we only publish the data of the group that was exposed to randomized arguments to ensure the data is not biased. The original statements are in German, but an English translation is supplied for better understanding of the dataset.

To get a feeling of how the data looks like, we describe the T_0 data (which is not part of any test set): There are 264 user profiles. In the context of the positions, 81% of the users support the prohibition of plastic packaging, 74% are in favor of the prohibition of genetic engineering. For the plastic topic, all pro-prohibition arguments are considered convincing by 81%; for genetic engineering, the number is 67%. The arguments against prohibition are convincing for 36%, or 41%, respectively.

The average length of the arguments in the initial argumentation pool compiled by us is 15.7 words (standard deviation 4.7). The mean length of the users' arguments is 10.4 words (standard deviation 7.3).

In the dataset provided, the user profiles are stored as a sparse matrix. The matrix for T_0 has 264 rows and 151 columns, of which at least 31 have a value (user name, opinion and strength on 2 positions, and at least 7 arguments per position with conviction and weight). The matrix for T_1 comprises all the user profiles from T_0 and, in addition, the profiles of new users from T_1, resulting in a matrix with 674 rows ($264 + 410$ users), and 151 columns. For T_2, the matrix contains a subset of updated rows of T_1; the users at T_2 are a subset of the T_1 users, i.e., users who left the empirical study between T_1 and T_2 are removed, leaving 553 rows; as new arguments were added, the matrix has 407 columns. Analogously, the matrix for T_3 is an update of the T_2 matrix and comprises 493 rows and 495 columns (note that there are not opinions for all statements, but only for a total of 247, as statements added by users from other experimental groups are also included).

4 Challenges and Baseline Results for Recommender Systems

Based on our dataset, we introduce three different classification and recommendation tasks where the opinions on statements and weights of arguments have to be predicted. We provide baseline results from a majority classifier and a neighbor-based recommendation algorithm to get a first feeling for the hardness of the tasks.

[2] `https://github.com/hhucn/argumentation-attitude-dataset`

4.1 Challenges

We propose the following three tasks on our dataset to show its applicability for further research on argument recommender systems:

1. Predicting a user's conviction
2. Predicting the argument weights
3. Predicting the most convincing arguments

For each task, it is possible to predict data from T_2 (for the *plastic packaging* topic) based on the data known at T_1 (i.e., including the data from T_0, which solves the cold-start problem), as well as the data from T_3 (*genetic engineering*) based on T_2. We will refer to those variants as $T_1 \to T_2$, or $T_2 \to T_3$, respectively. For dealing with sparse data, we follow an approach mentioned by Herlocker et al. (2004) for all tasks: We "ignore recommendations for items for which there are no ratings." The set of statements we evaluate a user $u \in U$ on with this approach is denoted as S_u. All prediction tasks are described in detail in the following.

4.1.1 Prediction of Conviction (PoC)

Based on the given data at time point T_i, predict whether the user considers an argument convincing (1) or not (0) for each user and each premise statement which was provided by us and for which the user opinion is known at time point T_{i+1}. The evaluation measure for this task is the mean accuracy: The accuracy for each user is calculated and then averaged over all users.

$$acc = \frac{\sum_{u \in U} \frac{\sum_{s \in S_u} [c(s) = \hat{c}(s)]}{|S_u|}}{|U|} \tag{1}$$

This tasks tests how good an algorithm can predict whether a user considers an argument the user has not seen before convincing.

4.1.2 Prediction of Weight (PoW)

Based on the given data at time point T_i, predict the weight for an argument (value in the interval $[0, 6]$) for each user and each argument which was provided by us and the user's weight is known for at time point T_{i+1}. We use the averaged root mean squared error as evaluation measure. This way, algorithms which produce some very bad predictions are punished.

$$rmse = \frac{\sum_{u \in U} \sqrt{\frac{\sum_{s \in S_u} \left(w(s) - \hat{w}(s)\right)^2}{|S_u|}}}{|U|} \tag{2}$$

Algorithms which perform well on this task are able to select arguments which are better suited to convince users.

4.1.3 Prediction of Statements (PoS)

Based on the given data at time point T_i, predict up to three statements the user considers convincing for each user and each premise statement which was provided by us and the user opinion is known for at time point T_{i+1}. We evaluate the macro precision on the created set of recommendations S_{u3} (which is commonly referred to as precision@3 (Silveira et al., 2019)).

$$p@3 = \frac{\sum_{u \in U} \frac{\sum_{s \in S_{u3}} [c(s) = \hat{c}(s)]}{|S_{u3}|}}{|U|} \tag{3}$$

In case S_{u3} is empty, that user is skipped in the evaluation. The goal of this task is measuring the quality of an algorithm's top recommendations, i.e., cases in which the algorithm is very sure that the user is convinced of a statement.

Many other tasks, e.g., predicting the opinion on positions, could also be looked at, but we limit ourselves to those three tasks in this paper. We think that the proposed tasks are important for applications which want to suggest interesting or persuasive arguments to a user.

Our dataset contains appropriate training data for the tasks we propose above, as well as a validate–test split (50%/50%): For each of the variants $T_1 \to T_2$, and $T_2 \to T_3$, the training data comprises the user profiles known at the points of time T_1, or T_2, respectively. The validation and test data contain the data of participants at T_2, or T_3, respectively, randomly assigned to either the validation or test dataset.

4.2 Baseline Results

We provide baseline results from a simple majority classifier and a more sophisticated nearest-neighbor (NN) classifier. The majority classifier always predicts the most common opinion of all users for which the opinion to be predicted is known (PoC) or considers the averaged weight (PoW and PoS).

The NN classifier was also used in our original research study to predict arguments that the users would most likely find convincing. We used it in some experimental groups, whereas other groups were confronted with randomly chosen arguments. We originally chose that algorithm on a best-guess basis because of a lack of suitable evaluation data

Table 3: Searched hyperparameter space.

n:	$5, 10, 20, 30, 40, 50, 100, 500$
α:	$0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$
depth:	$1, 2$

for comparing different algorithms before carrying out our study. Using our dataset, we can now quantify how good that algorithm actually is. By publishing our results we want to motivate other researchers to outperform our baseline results, and we provide an evaluation data set for future experiments that are similar to our own experiment.

The NN classifier uses the collaborative-filtering based recommendation algorithm from our argumentation tool *deliberate* (Brenneis and Mauve, 2020). To predict a value v, it first determines the n nearest users for whom the value to predict is known, using our pseudometric for weighted argumentation graphs (Brenneis et al., 2020). The pseudometric considers the attitudes of users and gives a higher weight to attitudes closer to the root of an argumentation (depending on a parameter α, where a lower α emphasizes positions over deeper statements in the argumentation tree, similar to the PageRank algorithm (Page et al., 1999)). Then, the value v of those nearest users is averaged, weighted by the calculated distance to each user.

The values for the hyperparameters have been chosen based on the results on the validation set. The search space is depicted in Table 3; all possible combinations were evaluated. The parametrizations used for each task are presented in Table 4.

Table 5 depicts the results on the test sets for both algorithms. From the results we can see that the NN algorithm performs better for all tasks and dataset combinations. The difference for the $T_2 \to T_3$ variant is always bigger than the difference for $T_1 \to T_2$. In the following section the results are discussed and analyzed in further detail.

The code to reproduce our results is provided together with our dataset.

4.3 Discussion of Baseline Results & Evaluation

From the increasingly greater difference of the NN algorithm, compared to the majority algorithm from $T_2 \to T_3$ to $T_1 \to T_2$, we can anticipate an NN algorithm to perform better on all tasks, if more thorough user profiles are available (remember that only two data points are known for participants in T_1). On the other hand, the description of our T_0 data has also shown that the arguments related to *genetic engineering* are considered less convincing on average than those for/against *plastic packaging*; this might be a disadvantage for the majority classifier when predicting the *genetic engineering* data for $T_2 \to T_3$. This could also explain why both algorithms perform worse when evaluated on data from T_3.

Although the NN approach outperforms the majority classifier, the difference is still quite small. It is certainly possible to build better predictors, maybe incorporating linguistic information of the arguments, e.g., the appearance of certain keywords, for instance “nature.” Another approach would be using different metrics for the NN classifier or applying a completely different machine learning method, e.g., decision trees or neural networks.

We chose evaluation measures which seemed sensible for us in our applications contexts, i.e., within the use case of the software *deliberate*. But depending on the application, other evaluation measures might be more sensible, like utility and novelty (Silveira et al., 2019), which might need more data on how a user consumed an argument (comparable to the click-through rate for search engine results).

The way we handled sparse data for the evaluation can also be discussed. Herlocker et al. (2004), who suggested “to ignore recommendations for items for which there are no ratings” for sparse data, also point out a disadvantage of this method, namely “that the quality of the items that the user would actually see may never be measured.” We do not think that this is a big issue in our evaluation context, since we basically evaluate the system on six randomly selected items per user for which the ratings are known.

5 Related Work

Similar datasets have been published before, and similar recommender tasks have been considered.

Habernal and Gurevych (2016) suggested the task of predicting convincingness of web argument pairs. They annotated and published a large-scale dataset of 16k argument pairs on 32 topics for the task of convincingness prediction and argument ranking. Different from our work, the task was not predicting the attitudes for each user for a given argument, but compare arguments in pairs and de-

Table 4: Hyperparameters for the nearest-neighbor classifier for each task, determined with the validation sets.

Task	n	α	depth of statements considered
PoC	20	0.5	2
PoW	100	0.5	1
PoS	10	0.5	2

Table 5: Results of our baseline methods on the test sets for the three different tasks for each dataset combination. NN always outperforms Majority.

Task	PoC (acc)		PoW ($rmse$)		PoS ($p@3$)	
Algorithm	$T_1 \to T_2$	$T_2 \to T_3$	$T_1 \to T_2$	$T_2 \to T_3$	$T_1 \to T_2$	$T_2 \to T_3$
Majority	.793	.639	1.80	1.95	.846	.627
NN	**.804**	**.675**	**1.74**	**1.82**	**.856**	**.677**

termine their objective convincingness.

Rahman et al. (2019) presented a dataset with 16 positions on 4 issues, for which 309 students gave their attitudes by adding arguments and indicating their level of agreement with that argument on a scale from -1 (total disagreement) to 1 (total agreement). Using the information about argument agreement, the agreement with the position was calculated. In our work, however, we explicitly ask for the agreement with a position, which allows a user to have an opinion which is inconsistent with their arguments. The authors also compared different algorithms for predicting user opinions on positions, where the best algorithm was a kind of soft cosine measure, which exploited feature similarity using position correlation.

Rosenfeld and Kraus (2016) tested different recommender agents in laboratory argumentation settings where arguments probably used next in a discussion were suggested. Different features were considered, i.a., the distance of arguments in the argumentation graph, a calculated argument strength, and the current context in the discussion. Several machine learning algorithms like SVMs and neural networks were evaluated. This is different from our work because we only recommend statements which are a premise for a given statement, although considering a broader suggestion strategy, which suggests statements from a different context, might be more appropriate for specific applications.

Chalaguine and Hunter (2020) presented a chat bot which should select appropriate counter-arguments, using cosine and concern similarity, with the goal of persuading a human to change their opinion. They compared their algorithms with a random baseline and got significantly better-than-random results for selecting relevant arguments. A crowd-sourced dataset with arguments about UK university fees was used (Chalaguine and Hunter, 2019). In contrast to our work, this dataset only contains arguments, but no user profiles with the attitudes of different persons on the arguments. The same applies to other corpora, like the Internet Argument Corpus (Walker et al., 2012).

6 Conclusion and Future Work

In our work, we introduce an extensive dataset which contains more than 900 arguments for two political positions and the user attitude data from more than 600 individuals, collected at different points of time. This dataset can be used for evaluating argument recommender systems, which can, e.g., be used to help people finding personally relevant arguments in discussions with many arguments. We suggest three different recommender tasks and provide baseline results from a simple majority predictor and a more sophisticated nearest-neighbor algorithm, which yields better results.

Our baseline results can still be improved on, and we invite everyone to develop better algorithms. Possible first improvements are considering linguistic information, and using different metrics for the nearest-neighbor classifier. What is more, other tasks could be defined on our dataset, e.g., predicting T_3 data from T_1 or non-convincing arguments. Furthermore, we want to research the effects of different recommendation strategies for argumentation on the formation of opinion when they are used to pre-filter content a user can see. Other evaluations in terms of novelty and utility should also be considered in the future.

Acknowledgments

We thank Marc Feger for translating the dataset. This publication has been created in the context of the Manchot research group *Decision-making with the help of Artificial Intelligence*, use case politics.

References

Trevor Bench-Capon and Sanjay Modgil. 2009. Case law in extended argumentation frameworks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 118–127.

Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems*, 46:109–132.

Markus Brenneis, Maike Behrendt, Stefan Harmeling, and Martin Mauve. 2020. How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs. In *Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020)*, number 2672 in CEUR Workshop Proceedings, pages 2–13, Aachen.

Markus Brenneis and Martin Mauve. 2020. deliberate – Online Argumentation with Collaborative Filtering. In *Computational Models of Argument*, volume 326, page 453–454. IOS Press.

Lisa Andreevna Chalaguine and Anthony Hunter. 2019. Knowledge acquisition and corpus for argumentation-based chatbots. In *CEUR Workshop Proceedings*, volume 2528, pages 1–14. CEUR Workshop Proceedings.

Lisa Andreevna Chalaguine and Anthony Hunter. 2020. A persuasive chatbot using a crowd-sourced argument graph and concerns. *Frontiers in Artificial Intelligence and Applications*, 326(Computational Models of Argument):9–20.

Carlos Chesnevar, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, Steven Willmott, et al. 2006. Towards an argument interchange format. *The knowledge engineering review*, 21(4):293–316.

Joe Collenette, Katie Atkinson, and Trevor Bench-Capon. 2020. An explainable approach to deducing outcomes in european court of human rights cases using adfs. *Frontiers in Artificial Intelligence and Applications*, 326:21–32.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.

Werner Kunz and Horst W. J. Rittel. 1970. *Issues as elements of information systems*, volume 131. Citeseer.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Md Mahfuzer Rahman, Joseph Sirrianni, Xiaoqing (Frank) Liu, and Douglas Adams. 2019. Predicting opinions across multiple issues in large scale cyber argumentation using collaborative filtering and viewpoint correlation. *The Ninth International Conference on Social Media Technologies, Communication, and Informatics*, pages 45–51.

Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4):1–33.

J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer.

Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. 2019. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10(5):813–831.

Statistisches Bundesamt (Destatis). Gesellschaft und Umwelt.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, volume 12, pages 812–817. Istanbul.

4.2 Building a Voting Advice Application Based on Arguments

Whenever a general election approaches, many voters start wondering which party to vote for. Different parties have different attitudes and arguments on political issues. *Voting Advice Applications* (VAAs) like the German Wahl-O-Mat try to assist voters with finding the political party which best matches their opinions. But those "classical" VAAs do not consider the reasons *why* parties and voters have certain views.

Therefore, we have built ArgVote, a new argument-based VAA. In this section, we first present the general idea of and motivation behind building ArgVote. We examine the results of an empirical study involving ArgVote. Among other findings, we discovered that providing arguments improves the understanding of different views and people enjoy interacting with an argument-based VAA. Subsequently, we look at the main decisions made during the development, e.g. how the argument corpus has been created. We study the user feedback we got and how it can be addressed in future work. Last of all, we propose a completely different interaction model for the VAA, namely an interactive chat bot.

4.2.1 Introduction of ArgVote

For a first introduction of the concepts in ArgVote and the reasons for building that application, we have a look at the peer-reviewed conference paper by Brenneis and Mauve (2021b), in which ArgVote has been presented for the first time:

Markus Brenneis and Martin Mauve.

"ArgVote: Which Party Argues Like Me? Exploring an Argument-Based Voting Advice Application"

Proceedings of the 13th KES-IDT 2021 Conference, *Intelligent Decision Technologies*, pages 3–13, Springer Singapore.[2]

With this peer-reviewed paper, we contributed the following:

1. motivation for an argument-based VAA
2. introduction of ArgVote, a VAA which considers opinions on arguments in its matching algorithm
3. design of an empirical study with ArgVote regarding the influence of arguments within an VAA on informedness, ease of indicating an opinion, ease of use, and party–voter matching

[2]Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Intelligent Decision Technologies, ArgVote: Which Party Argues Like Me? Exploring an Argument-Based Voting Advice Application, Markus Brenneis and Martin Mauve, 2021.

4. findings from that study
5. an argumentation dataset based on the Wahl-O-Mat for the European Parliament Election 2019 for the six biggest parties, and a dataset with user profiles of 30 study participants

Datasets from VAAs are a valuable resource since they are already used by "political analysts, social scientists, political parties or any other independent organization in order to discover knowledge about the electorate's perceptions and feelings on certain issues, voting behavior, relationships between voters and candidates as well as about many other issues" (Katakis et al., 2013). For example, political parties could learn where they could make small changes to their policies to extend their electorate. Therefore, an extended dataset which also contains argument information is even more valuable. One could imagine the automated finding of compromises or even a kind of "digital deputy" that reflects the opinions and arguments of all users (although it is probably hard to find compromises for conflicting goals).

Personal Contribution

Martin Mauve had the original idea of building a VAA which considers arguments in its matching algorithm. The design of the VAA was developed by Markus Brenneis, the author of this thesis. Stefan Harmeling came up with the UI idea that arguments should be within an expandable view ("expert view"). Markus Brenneis wrote the software, developed the initial experiment design, conduced the survey, and wrote the whole paper. His questionnaire design was discussed with Henrik Domansky, Stefan Marschall, Martin Mauve, and Lucas Constantin Wurthmann. Martin Mauve provided feedback on drafts of the paper. Markus Brenneis, Marc Feger, and Jan Steimann jointly created the party attitude dataset used for the evaluation of the software.

Importance and Impact on This Thesis

ArgVote is the second user-facing application of our pseudometric from Section 3.2. In this application context, the pseudometric is straight-forwardly used to determine voter–party similarity. A justification that our pseudometric is able to yield intuitive results is given by our work from Sections 3.3 and 3.4. The empirical study conducted with ArgVote gives important insight into whether the pseudometric is accepted and works well in an application context. Some design aspects of ArgVote are based on our findings with *deliberate*, which were discussed in Subsection 4.1.4.

ArgVote: Which Party Argues Like Me? Exploring an Argument-Based Voting Advice Application

Markus Brenneis and Martin Mauve

Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany,
Markus.Brenneis@uni-duesseldorf.de,
https://cs.hhu.de/

Abstract. A lof of people use Voting Advice Applications (VAAs) as a decision-making tool to assist them in deciding which political party to vote for in an election. We think that arguments for/against political positions also play an important role in this decision process, but they are not considered in classical VAAs. Therefore, we introduce a new kind of VAA, *ArgVote*, which considers opinions on arguments when calculating voter–party similarity. We present the results of an empirical study comprising two groups who used ArgVote with and without arguments. Our results indicate that arguments improve the understanding of political issues and different opinions, and that people enjoy the interaction with arguments. On the other hand, the matching algorithm which considers arguments was not better, and user interface improvements are needed. The user profiles we collected are provided to assist further research.

Keywords: Argumentation, Data Set, Voting Advice Applications

1 Introduction

Many people [1, 2] around the world use voting advice applications (VAA) like *Vote Compass* or the German *Wahl-O-Mat*. They inform themselves about positions of different parties concerning current political issues before general elections to receive help in deciding for whom to vote. In many applications, the similarities between voters and parties are calculated with a high-dimensional proximity model [3], based on proximity voting logic [4], where parties are matched with voters based on their opinions concerning a number of political positions.

Classical VAAs, however, do not consider *why* parties and voters maintain certain views. Consider, for instance, Party A being against nuclear power because it thinks nuclear power plants are dangerous, and Party B is against nuclear power because nuclear waste cannot be stored safely. If a voter thinks that nuclear power plants are safe, they are certainly closer to Party B than to Party A. But a classical VAA, which only asks whether the voter is for or against nuclear power, would not capture this information. Therefore, we assume that not only the opinions concerning political positions, but also the arguments used to sustain these positions are relevant for the personal party preference.

The final authenticated version is available online at https://doi.org/10.1007/978-981-16-2765-1_1.

Hence, we have developed *ArgVote*, a new kind of VAA, which does not only consider the political positions, but also the arguments used to arrive at the given position. In an online survey comprising two groups, we tested the acceptance of our new application and whether its new matching algorithm performs better than that of a classical VAA. We also questioned whether people are more informed when arguments are presented, and if they can indicate their own political opinion more easily.

In the next section, we explain why and how we developed an argument-based VAA. Then we present our methods and hypotheses. In Section 4, we show our results and subsequently discuss their consequences. Finally, we have a look at related work and summarize our findings.

2 Designing an Argument-Based VAA

We now sum up the key motivations for developing an argument-based VAA, and then present how our new application ArgVote looks like.

2.1 Limitations of Classical VAAs

As described in the introduction, we think that the reasons why a party has certain attitudes are also important for providing sensible support for a voting decision. If, in our example, the problem with nuclear waste was solved, then Party B would be likely to change its attitude towards nuclear power, as would a voter who was against nuclear power for the same reason. This reinforces our stance that arguments are relevant.

What is more, voters might not be familiar with an issues raised within a VAA, and they tend not to "look up additional information on the web and oftentimes 'just' provide a neutral no opinion answer" [5]. We conjecture that providing arguments for and against a position right within the VAA increases the informedness of voters, who can then better express their opinion and get more meaningful results, i.e. a more suitable voting advice.

Another advantage of arguments is making it harder for parties to "cheat" when the parties provide the answers to the questions in the VAA themselves. Sometimes, parties indicate to be neutral instead of taking an unpopular position to improve their results [2], which leads to inconsistencies between the official stance of a party and its reasons.

2.2 How ArgVote Works

Based on the design of the German VAA Wahl-O-Mat, we have developed ArgVote which additionally displays arguments for and against agreeing with a position (see Figure 1). The arguments can be displayed before the voter indicates their opinion, but ArgVote also explicitly asks the voter to (optionally) choose their arguments after opinion input. If available, (counter)arguments for/again the arguments displayed can be navigated through. The arguments

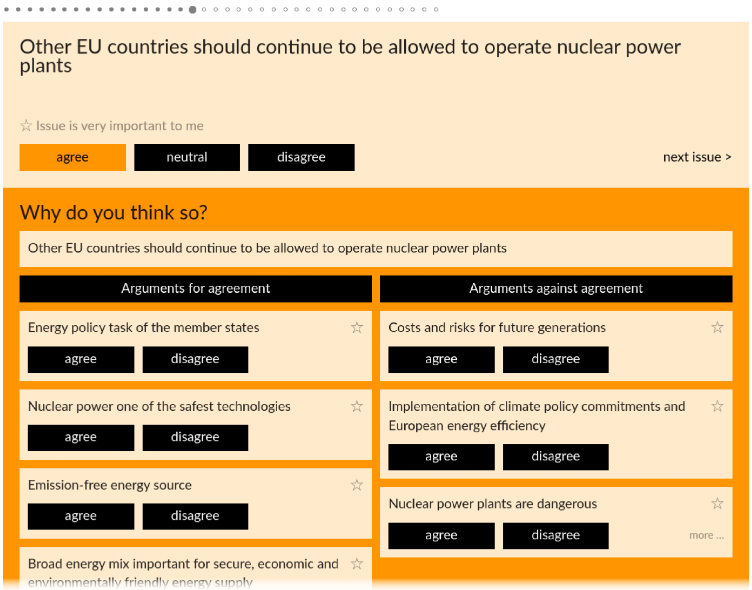


Fig. 1. Main user interface of ArgVote: The user is asked for their arguments after indicating their opinion on an issue, but they can also display the arguments beforehand. "More ..." can be used to see (counter)arguments to arguments.

presented in ArgVote were provided by the parties beforehand. As in the Wahl-O-Mat, political issues, but also arguments, can be marked as important, giving them a higher weight in the matching algorithm. After the last question, a user can compare their arguments with the parties' arguments and sees a bar chart indicating how much they agree with the attitudes of the individual parties.

In the classical matching algorithm used by the Wahl-O-Mat [6], party and voter have a distance 0 for an issue if they have the same opinion, 0.5 if they are different and one is neutral, and 1 otherwise; the value is doubled for issues marked as *important*. ArgVote's matching algorithm is based on our pseudometric for weighted argumentation graphs [7], which also considers the opinions on arguments for/against the positions. *agree*, *neutral* and *disagree* are translated to opinion values 0.5, 0, or -0.5, respectively, in this model. When a position or argument is marked as important, the corresponding edge in the argumentation tree gets a doubled weight.

The relative importance of opinions on arguments and positions can be balanced with a parameter α of the used pseudometric (similar to PageRank's [8] damping factor). ArgVote uses $\alpha = 0.3$, giving the opinions on positions a slighter higher influence than the arguments used. This choice is motivated by the results of an earlier empirical study of ours [9], which indicated that opinions on positions are considered more important by most people than opinions on arguments. From the same study, we also learned that the results of the chosen pseudometric matches human intuition well, and thus, are understandable, in many argumentative contexts.

3 Hypotheses and Methods

With ArgVote, we want to identify differences after using an argument-based VAA and a classical VAA. For our experiment, we recruited German participants from within our personal contacts[1] and let them use ArgVote in two different modes: Group 1 used ArgVote as described above, Group 2 (control group) used ArgVote without arguments displayed under the theses, i.e. it basically behaved like the Wahl-O-Mat. Before the participants used ArgVote, we asked them for their sympathy with the biggest German parties (in alphabetical order: AfD, CDU, Die Linke, FDP, SPD, and Grüne), which were included in ArgVote.

The content for ArgVote was copied from the Wahl-O-Mat of the European Parliament Election 2019, which had been the last election where all Germans were allowed to vote, and comprised 38 positions. We only used the first 15 positions in both groups to reduce the time needed for participation. The complete argumentation corpus contains 294 arguments for all political theses and 147 arguments for the first 15 issues. It was created by three annotators based on the justification statements the parties provided in the Wahl-O-Mat. All annotators independently annotated for each argument whether it is used by a party. The annotator agreement in terms of Krippendorff's alpha [10, p. 211 ff.] is 78%.

In our experiment, we want to research the differences between both groups regarding subjective informedness, ease of indicating an opinion on a thesis, better matching results compared to own party preferences, and usability assessment of ArgVote. After using ArgVote, we asked participants what features of ArgVote they used, how hard they were to use, and how well-informed they feel about policies. Moreover, we count how often user indicate no opinion.

ArgVote also asks different questions about how much participants like their matching results (in overall and concerning the top position) to get a subjective rating of how good the result is. We also checked how close the calculated matching matches a participant's party sympathy rating using the rank-biased overlap (RBO) [11]; RBO compares two sorted lists, where difference in the top-positions are punished more than differences in bottom-positions. We also compare the average rank of a user's party, as also done before for other VAAs by [3].

To wrap up, we have the following hypotheses:

1. Group 1 feels more informed after using ArgVote with arguments than Group 2 (control group without arguments).
2. It is easier for Group 1 to indicate an opinion for a political thesis.
3. Group 1 does not consider ArgVote harder to use.
4. Matching results of Group 1 better match participants' party preferences.

We want to clarify that we mainly focus on checking whether our general idea works well. If it works well, a bigger study can be considered, where improvements on the user interface, the selection and formulation of the arguments, and a more representative sample can be considered.

[1] We first planned to do on-campus recruiting of participants, but this was not possible due to the lockdown at that time.

4 Results

We now present our key findings, starting with the comparison of the experimental groups, and then checking our hypotheses presented in the previous section. The dataset containing the VAA questions and argumentation corpus, as well as the collected user profiles are provided online[2].

4.1 General Information on Participants and Groups

60 participants successfully completed our survey (including two attention check questions). 30 were in Group 1 (with arguments), 30 in Group 2 (control group without arguments). 63% of the participants were male (German population: 49% [12]), the average age was 27 (German population average: 45 [12]), and more than 96% had at least a higher education entrance qualification (Hochschulreife; German average: 34% [12]).

4.2 Hypothesis 1: Informedness

Looking at the subjective answers about informdness, which had been asked after using ArgVote and are presented in Table 1, we could not deduce that Group 1 got a higher awareness of political topics, nor the differences of parties became clearer. But we saw that Group 1 got a clearer picture why there were different opinions, and they understood political issues significantly better.

Table 1. Subjective level of informdness on a Likert scale from *do not agree at all* (1) to *fully agree* (5), p-values according to a Mann–Whitney rank test (MW) [13].

Question	Group 1	Group 2	p (MW)
By using ArgVote I became aware of political issues.	**2.87**	2.60	.20
After using ArgVote, the difference between the parties is clearer to me.	2.70	**2.87**	.77
Using ArgVote helped me understand some political issues better.	**3.40**	2.33	$< .001$
After using ArgVote, it is clearer to me why there are different opinions on certain theses.	**3.30**	2.77	.054

4.3 Hypothesis 2: Ease of Indicating Opinion

There was no big difference between both groups regarding the number of neutral answers and skipped questions. On average, 28% of participants in Group 1 chose a neutral answer, whereas 30% of participants in the control group did so (no significant difference, $p = 1$ with a χ^2 test). The average skip rate (i.e. providing no opinion on an issue) was 1.1% in Group 1, and 1.3% in Group 2 ($p = .58$).

[2] https://github.com/hhucn/argvote-dataset

Table 2. Assessments of task difficulty on a Likert scale from *very hard* (1) to *very easy* (5).

Question	Group 1	Group 2	p (MW)
give my opinion on the theses	**3.67**	3.47	.25
mark a thesis as important	**3.32**	3.24	.36
agree/disagree with arguments	3.78	n/a	n/a
mark an argument as important	3.48	n/a	n/a

On the other hand, the subjects in Group 1 considered indicating an opinion on a thesis slightly easier than those in Group 2 (cf. Table 2). We can also see that (dis)agreeing with arguments was not considered much more difficult than giving an opinion on a thesis, which means that this additional task was not too hard for VAA users. Group 1 also strongly agreed that seeing arguments next to the theses is useful (4.40 on a Likert scale from 1 to 5).

4.4 Hypothesis 3: Ease of Use

As depicted in Table 3, subjects in both groups understood ArgVote, had no problems with navigating, and tended to use the tool again. Group 1 considered the user interface more cluttered and less self-explanatory, which makes sense because of the additional features available. Surprisingly, Group 1 enjoyed using ArgVote more, maybe because it offered a new kind of interaction.

Table 3. Assessments of usability on a Likert scale from *do not agree at all* (1) to *fully agree* (5).

Question	Group 1	Group 2	p (MW)
ArgVote appeared cluttered to me.	2.23	**1.87**	.90
ArgVote was self-explanatory.	3.83	**4.26**	.94
I did not understand how ArgVote works.	1.47	**1.30**	.75
I had no problems navigating ArgVote.	4.16	**4.33**	.69
I would use ArgVote again.	4.20	**4.23**	.63
I enjoyed using ArgVote.	**3.93**	3.60	.052

On the objective side, the time participants stayed in ArgVote and its introduction page was significantly longer in Group 1 (median 17.9 minutes) than in Group 2 (6.08 minutes). This increase was expected because interacting with the arguments needs more time, but it also shows that participants actually did spend time with arguments and did not ignore them.

4.5 Hypothesis 4: Better Matching

We anticipated that taking into account the opinion on arguments (Group 1) yields results better matching individuals' party preferences. In fact, the RBO (with its parameter $p = 0.7$) in Group 1 (0.67) was worse than the RBO for Group 2 (0.71). Looking at how often the calculated top-1 position matches the party preference, we see something similar (Group 1: 37%, Group 2: 50%). We also got better results for Group 2 when considering the average position at which the user's preferred party is put (Group 1: 1.97, Group 2: 1.60).

Table 4. Assessments of the party matching after using ArgVote on a Likert scale from *do not agree at all* (1) to *fully agree* (5).

Question	Group 1	Group 2	p (MW)
I am confused about the result.	**2.07**	2.17	.35
I am happy with which party is displayed at position 1.	3.80	**3.87**	.62
I can understand which party is displayed at position 1.	**4.23**	4.10	.41
I can understand the displayed percentage of agreement with the party at position 1.	4.00	**4.07**	.75
I consider the overall order of the parties as a whole to be reasonable.	3.90	**3.97**	.63
I consider the percentage of agreement of the parties as a whole to be reasonable.	**3.97**	3.60	.12

The subjective satisfaction with the matching result was basically the same in both groups (cf. Table 4). Participants in Group 1 understood the percentages presented in the matching slightly better.

5 Discussion

Our results give a first hint that incorporating arguments in a VAA makes sense since people tended to be more informed, to give their own opinion more easily and enjoyed the new kind of interaction. But there are some limitations in our current approach, especially when considering using ArgVote for a real election.

We are well aware that our participants were not representative for the German population, which was due to our recruiting process which mainly targeted young students at a university. But our results still give first important hints on whether our approach of incorporating arguments into a VAA is sensible. A bigger study with older and less educated people would be needed, though, to see if they perceive ArgVote as positively as our highly academic, young sample.

The user interface (UI) was considered more cluttered, hence reducing the pieces of information shown at once should be considered, e.g. by pre-filtering the arguments presented in a sensible way. This could, however, lead to the feeling of being manipulated. Related to this, a mobile-friendly UI is not yet available, but is important in a time in which most site views on the Internet come from mobile devices.

From free text comments, we could also learn that the UI regarding the presentation of arguments should be improved, e.g. it was not always clear what "agreeing" with an argument means ("the argument makes sense to me in this context" vs. "in my view, this statement is correct (but is possibly no good argument)"). Some users also wished to "partially agree" with theses or arguments, as possible e.g. in the VAA ParteiNavi; this could be handled by the underlying pseudometric, but was not possible through our UI.

An important question is where the argument come from. For simplicity, we used the arguments provided in the parties' statements in the Wahl-O-Mat for each issue in our experiment. It can be assumed, though, that a party does not mention every argument it (dis-)agrees with in its statement, which means that the dataset created that way is incomplete. Furthermore, the general party sympathy might also not be in line with the stance on the 15 European topics presented to the participants. Those aspects could also explain why the argumentation-based matching algorithm performed worse when compared to party sympathy, but had a better subjective rating.

A better approach would be asking parties to provide all their arguments and also providing opinions on other parties' arguments, possibly through argumentation platforms like kialo. It has to be considered, though, that participation in such a platform would be hard for small parties with few resources. A related question is whether voters should be able to provide arguments for the corpus, too. Furthermore, some parties mention compromise proposals in their reasons for positions, but those cannot be mapped to arguments, and hence, cannot currently be presented in ArgVote.

Another aspect is whether reaching a good agreement with a user's party preference and user satisfaction are actually the goal of a VAA. The personally preferred party might actually not match the party which would represent one's interests best, but lacking a sensible ground truth, we think that party sympathy is the best approximation we can get.

6 Related Work

We are not aware of other VAAs which incorporate opinions on arguments in their matching algorithms. But there are other kinds of VAAs which also use other approaches than pre-defined distance functions to determine party–voter similarity of classical VAAs, or provide arguments in their interface.

So-called Social VAAs (SVAAs, e.g. *Choose4Greece*) use collaborative filtering, where recommendations are made based on the voting intention of similar users.[14] As shown in [15], the results of SVAAs can be better than those of traditional VAAs. For evaluation purposes, the voting intention given by the users was used, which has the limitations we have already discussed in Section 5. A problem of model-based SVAAs is that their results are not easy to explain[4], whereas understandable results were a design-goal of *ArgVote*, which influenced the choice and design of the underlying pseudometric.

The Learning VAA by [4] took another approach by learning individual distance matrices for each issue instead of using one global, fixed distance function.

Finding political parties closest to a user was also studied in [16]. A user's party could be predicted based on their opinions on ideological positions with an accuracy of 80%. They used a dataset from debate.org, and applied collaborative filtering to make clustering with sparse information easier. It has to be noted, though, that there have only been two parties (Republicans and Democrats) in that experiment, but we considered six parties.

Some VAAs, like the Greek *Votematch* or *VoteSwiper*, can show additional information on a position, including arguments, but they do not consider argument agreements in their matching. Similarly, the Dutch VAA *Young Voice* provided short videos with pro and contra arguments for each thesis. In the study presented in [17], showing additional information like arguments did increase the number of issues for which an opinion was given, but it did not improve the comprehension of issues. We could not confirm this result in our study.

Our argument annotation process was similar to the method by [18]. They first created a corpus of all possible arguments, and then multiple annotators decided for each text–argument pair if the argument is present in the text.

7 Conclusion and Future Work

We have introduced ArgVote, a new kind of Voting Advice Application which can display arguments next to theses and considers opinions on arguments when calculating the user–party similarity. In an empirical study, we compared ArgVote with and without arguments. We got first hints that the arguments help with forming an opinion on a thesis, understanding different political positions, and make users enjoy the application. The matching results matched the subjective party sympathy worse, though, and our sample was not representative.

The dataset with arguments and user profiles is provided to the community, e.g. for improving the matching algorithm. Using this dataset, the performance of other, possibly more intelligent matching algorithms can be evaluated.

For future work, the user interface of ArgVote should be streamlined to feel less cluttered and reduce the time needed to use the VAA. One possibility would be considering completely different user interactions, e.g. an interactive chat bot, which could reduce the perceived time needed for dealing with the VAA by asking questions on different days. A more representative study should check the influence of the arguments on people who are older and less educated. Another major open question for a real-world application is how the argument corpus should be created.

Acknowledgements We thank the students Marc Feger and Jan Steimann for helping with the annotation of the argumentation dataset, and Henrik Domansky, Lucas Constantin Wurthmann, and Stefan Marschall for providing valuable feedback on the study design.

References

1. Van Camp, K., Lefevere, J., Walgrave, S.: The content and formulation of statements in voting advice applications. Matching Voters With Parties and Candidates. Voting Advice Applications in Comparative Perspective pp. 11–32 (2014)

2. Wagner, M., Ruusuvirta, O.: Matching voters to parties: Voting advice applications and models of party choice. Acta politica **47**(4), 400–422 (2012)
3. Mendez, F.: What's behind a matching algorithm? a critical assessment of how voting advice applications produce voting recommendations. Matching Voters with Parties and Candidates (2014)
4. Romero Moreno, G., Padilla, J., Chueca, E.: Learning VAA: A new method for matching users to parties in voting advice applications. Journal of Elections, Public Opinion and Parties pp. 1–19 (2020)
5. Kamoen, N., Holleman, B.: I don't get it: Response difficulties in answering political attitude statements in voting advice applications. In: 7th International European Survey Research Association Conference, vol. 11, pp. 125–140. European Survey Research Association (2017)
6. Marschall, S.: The online making of citizens: Wahl-O-Mat. The making of citizens in Europe: New perspectives on citizenship education pp. 137–141 (2008)
7. Brenneis, M., Behrendt, M., Harmeling, S., Mauve, M.: How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs. In: Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020), no. 2672 in CEUR Workshop Proceedings, pp. 2–13. Aachen (2020)
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (1999). Previous number = SIDL-WP-1999-0120
9. Brenneis, M., Mauve, M.: Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations. In: B. Fazzinga, F. Furfaro, F. Parisi (eds.) Proceedings of the Workshop on Advances In Argumentation In Artificial Intelligence 2020, no. 2777 in CEUR Workshop Proceedings, pp. 1–15. Aachen (2020). URL http://ceur-ws.org/Vol-2777/paper21.pdf
10. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, 2 edn. Sage Publications (2004)
11. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Transactions on Information Systems (TOIS) **28**(4), 1–38 (2010)
12. Statistisches Bundesamt (Destatis): Gesellschaft und Umwelt. URL https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/_inhalt.html
13. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics pp. 50–60 (1947)
14. Tsapatsoulis, N., Agathokleous, M., Djouvas, C., Mendez, F.: On the design of social voting recommendation applications. International Journal on Artificial Intelligence Tools **24**(03), 1550,009 (2015)
15. Katakis, I., Tsapatsoulis, N., Mendez, F., Triga, V., Djouvas, C.: Social voting advice applications - definitions, challenges, datasets and evaluation. IEEE Transactions on Cybernetics **44**(7), 1039–1052 (2013)
16. Gottipati, S., Qiu, M., Yang, L., Zhu, F., Jiang, J.: Predicting user's political party using ideological stances. In: International Conference on Social Informatics, pp. 177–191. Springer (2013)
17. Tziraki, C.M.: Towards accurate voting advice: How does Additional Information and Party Identification Influence the users' Comprehension of, and Answers to VAA Statements? Master's thesis, Tilburg University (2020)
18. Boltužić, F., Šnajder, J.: Back up your stance: Recognizing arguments in online discussions. In: Proceedings of the First Workshop on Argumentation Mining, pp. 49–58 (2014)

4.2.2 Major Decisions During the Development of ArgVote

While we were developing ArgVote, different questions regarding its design arose. Let us now have a closer look at how and why the argument corpus for our experiment has been created, which limitations we can identify in the argument-based approach, and how and why we used our model of weighted argumentation graphs and our pseudometric for calculating the voter–party similarity.

Source of Arguments

The first main issue was where the argumentative content comes from. In the Wahl-O-Mat, each party formulates its opinion independently of the others (Schultze, 2012). If we applied this to our argumentation context, this would mean that parties mainly put forward their own supporting arguments, but do not argue against other party's arguments. Furthermore, parties usually do not state which arguments they do not agree with. This leaves an incomplete view of the arguments for/against a thesis and deeper arguments as well as a party's attitude.

These problems could be addressed by letting all parties argue on an argumentation platform like *deliberate*, where they can react to other arguments and indicate their attitudes. A problem of this approach is that small parties do not have as many resources to participate in such a discussion as bigger parties, and parties would have to be willing to engage in the discussions. If there was such a platform, one could also ask whether voters should also be able to participate in the platform to bring in their thoughts.

In our first experiment with an argument-based VAA, we focused on checking whether the general idea works. Hence, we applied a simple method to gather an argumentation corpus by manually extracting arguments from existing party statements. But, as we will see in Subsection 4.2.3, some users negatively noticed that, in their opinions, some arguments are missing, have complicated wording, or are unevenly distributed between pro and contra, leading to possibly unintended influences on a user's opinion. Thus, if ArgVote should be used for a real election, special care must be taken when creating the argumentation dataset.

Limitations of the Argument-Based Approach

During the annotation process for the argument dataset creation, we discovered some limitations of the argument-based presentation in ArgVote. We decided to ignore them for our first experiment, but want to point out the consequences.

As already discussed above, extracting arguments from the parties' statements can lead to a heavily unevenly distribution arguments of between pro and contra. For example, the thesis about the introduction of same-sex marriage only has one supporting argument (provided by one party), but four counterarguments (from two parties); other parties simply did not provide argumentative content in their statements, although there are more supporting arguments which could be provided.

Another problem regarding the same issue is that the AfD mentions the counterargument "particular protection of the family in the Basic Law." This statement is objectively true since the Basic Law has a particular protection for marriage and family, so the statement could be agreed with. But it is disputable whether it is actually an argument against same-sex marriage, since the Federal Constitutional Court ruled that the special protection of marriage in the Basic Law does not prohibit introducing a similar concept for same-sex partnerships (BVerfG, 2002). So if a user agrees that the argument can be undercut, they would *not* agree with the argument. As we will see in Subsection 4.2.3, this is a real issue because some users were actually confused about what the agreement with an argument means. Moreover, a "complete" argument-based VAA should display the Court's ruling as a counterargument, but the argument is missing in the dataset because no party mentioned it.

We also noticed that party statements and opinions are sometimes inconsistent. For instance, the AfD agreed that "animal testing should still be allowed in medical research," but they only mention arguments against the thesis, e.g. "animals are fellow creatures and no objects." This can mean that the party forgot to actually back-up its stance, or it tried to "cheat" to get more matches with voters. In the latter case, considering the party's arguments in the matching algorithm, which ArgVote does, would mitigate the influence of cheating.

Some party statements also include non-argumentative content which cannot be mapped to ArgVote's argument-based design. For example, regarding whether the EU member states should build a common army, the SPD says that "the deployment of a European army may only be authorized by the European Parliament." This is a kind of conditional agreement, which is neither supported by ArgVote nor the Wahl-O-Mat. Adding support conditions and different interpretations of a thesis is an open question for future work and has also been suggested as part of the participants' individual feedback (cf. Subsection 4.2.3).

Possible Pre-Filtering of Arguments

For some theses, there are many arguments and it is time-consuming to read all of them. We could have decided to pre-filter the argument list, similar to what our software *deliberate* does. But it is unclear how the arguments can be selected transparently, without manipulating the user. Once enough user profiles have been collected, a collaborative filtering approach could be used to display arguments the user is likely to accept or decline, such that arguments with probably neutral stance are not displayed. To avoid the feeling of manipulation, users must always be able to see all arguments when they wish.

Collaborative filtering has the danger of building filter bubbles (Pariser, 2011), and "unscrupulous users" (O'Mahony et al., 2005) could theoretically try to manipulate the system such that it makes wrong predictions. To mitigate this problem, a hybrid-approach which also always displays random arguments could be considered. Another possibility is displaying only arguments for which the user's answer would actually change the matching result. For instance, if an argument for the last question in the VAA is only agreed to by the party which is currently ranked at the last place without the chance of getting a better rank, the argument could be hidden.

As we focused on studying the general influence of showing arguments, we decided to display all arguments without pre-filtering and not evaluate different filtering strategies. But as the UI of ArgVote was considered crowded by the users, a more compact display should be investigated in the future.

Mapping to Weighted Argumentation Trees

We now explain how we mapped party and user attitudes to our model of weighted argumentation graphs. When this mapping is done, the closeness of a user and a party can be calculated with our pseudometric (cf. Section 3.2). To get a sensible percentage value, the result of the function is normalized by the maximum possible value for a tree with depth 2.

All theses in the VAA are represented as positions. All edges from a thesis to the root node have the same weight, except if a thesis is marked as important. An important thesis has a doubled weight. When a thesis is agreed to, it gets a rating of 0.5, in case of disagreement -0.5, and otherwise 0.

A similar system applies to the arguments. If an argument presented is (dis)agreed to, the premise's node gets a rating of 0.5, or -0.5, respectively. Arguments without opinion on the premise have a weight of 0. As there are nearly no arguments on a deeper level, everything on a depth below level 2 is ignored.

Choice of Metric and Manipulability

In Section 3.4, we examined different metrics regarding matching human intuition in different argumentation scenarios. The argumentation hierarchy within ArgVote's model has at least two levels (positions and arguments for/against them). We personally anticipated and also confirmed in Section 3.3 that the opinions concerning positions are more important than the arguments used. Therefore, we needed a distance function which can balance the influence of positions and arguments, leaving us with the tree similarity measure by Bhavsar et al. (2004) and our own pseudometric. As the latter performed better in our comparison, we used it in ArgVote. From the comparison we also anticipated that our pseudometric would outperform the VAA distance, which did not hold and is analyzed further in Subsection 4.2.3.

An open question regarding the choice of the distance function is how hard it is to manipulate its results, as we already considered in Subsection 3.2.3. For example, could a party improve their ranking by leaving out unpopular arguments or agreeing with popular arguments which are against its position? One has to note, though, that the same problem is true for classical VAAs, where parties are known to sometimes cheat (Schultze, 2012; Wagner and Ruusuvirta, 2012).

4.2.3 Additional Findings and User Feedback

We now concentrate on further findings we got from our survey data, which we could not include in our paper (Brenneis and Mauve, 2021b). We start with additional insights from the questionnaire answers and the user profiles and then look at suggestions from users' free text comments.

Before the introduction of ArgVote, the participants tended, on average, to agree that a VAA should consider arguments in its matching algorithm (3.7 on a scale from 1 to 5). Being able to see arguments got an average rating of 3.1. After having used ArgVote, the usefulness of having seen the parties' arguments was rated with 4.6, which means that this feature was considered very important by most users.

We asked how well participants understood important political issues before and after having used ArgVote. Surprisingly, participants of both groups indicated, on average, a slightly worse understanding after using ArgVote, with a difference of -0.2 in both groups. The slight decrease can be explained with having been exposed to potentially unfamiliar political issues in ArgVote.

In our paper, we mentioned that the *Rank-Biased Overlap* (RBO) of party preference and ArgVote's rating was worse for Group 1 than for Group 2 (control group). This observation still holds when we only look at Group 1: If we calculate the RBO once with the classical VAA distance, and once with our pseudometric considering arguments, we get RBOs of 0.70, and 0.67, respectively. We also asked ourselves how often the order of parties provided by both distance functions is different, i.e. how often our pseudometric actually yields a different result. In fact, the order was the same only for 27% of the users, which means that it does not simply produce the same results as the classical matching algorithm for most users. In addition to the limitation of looking at the RBO mentioned in our paper, we have to take into account that we compare with the general party sympathy, which might not be relevant in the context of European questions. This is, however, a systematic error in both groups.

In our data, we can see that the pseudometric, compared with the VAA distance, produces rankings where the distance between the first and the last party are smaller (27 percent points, and 39 points, respectively), and the average matching percentage is also smaller (47, and 56, respectively). One user in Group 1 actually noticed this peculiarity. A possible explanation for the ranking becoming "closer" with our pseudometric is that any pair of argumentations probably overlaps and disagrees in some points. This means that a user becomes closer to parties they usually agree less with and further away from parties where there is an agreement on theses, but not on every argument.

Our pseudometric assumes that "neutral" is on a straight line between "agreement" and "disagreement," but we have seen in Subsection 3.2.2 that this assumption does not match human intuition when comparing argumentations. Nevertheless, we considered our pseudometric appropriate for use in the VAA in Subsections 3.4.3 and 4.2.2, and the same limitation is also present in the classical VAA distance. Testing the performance of another distance function taking this particularity into account would be interesting and can be done in future work using our dataset. We also treated neutral and no opinion/skip the same way, which is another limitation which could be prevented by ignoring skipped answers at least for theses. This solution could, however, lead to a loss of desired properties like fulfilling the triangle inequality, in particular if parties are also allowed to not answers questions.

More than 20 users provided feedback on ArgVote as free texts. We now summarize the main feedback which should be considered when developing improved versions of ArgVote.

As already touched on in the paper, several users asked for a possibility to partially (dis)agree with positions and arguments. One user suggested adding a slider, but explicit buttons "partially (dis)agree," as used by the VAA ParteiNavi, are probably more intuitive for most users. Presenting that possibility is certainly harder UI-wise, especially when considering the development of a mobile-friendly UI where less space is available. This improvement can, however, be directly handled by the underlying model using ratings.

Another point which was mentioned several times is the understanding of arguments. Many participants indicated difficulties with understanding the statements and asked for more clarifications, e.g. the contents of the refugee deal with Turkey, or explanations for terms like *decarbonisation*. Links to external resources or popups explaining statements in more detail could be added, but might, again, clutter the interface.

The current presentation of arguments was confusing for some users because they did not know what "agreeing" with an argument meant (i.e. whether the statement is accepted, or whether the statement is a good argument). Rephrasing the arguments with "... because" was suggested by a user. In the UPEKI project, "convincing" was used instead of "agree" in the context of arguments. What is more, navigating to arguments of arguments with the "more ... " button was confusing for a user, but most users who made use of this feature considered it useful (3.9 on a scale from 1 to 5, considering only 20 of 30 users who did use that feature). For comparison, the usefulness of marking an argument or thesis as important was rated with 3.1, and 3.7, respectively.

Participants of the control group were confused by the way the parties' arguments were displayed, which can be seen in Figure 4.1. The left/right division for pro/contra arguments were not clear for those users since they did not know this division from ArgVote's main view. As the control group view is not meant for a real application, this UI issue is not a big problem, but a clearer view should still be considered. Furthermore, the ticks indicating agreement with an argument were confusing, especially considering that parties seldom explicitly disagree with arguments. One user of Group 1 suggested an inverted view, where the arguments are the top items, and ArgVote should display next to each argument which parties agree with it. Another unclear aspect here was whether an empty argument list for a party is a bug, or if the party actually did not provide arguments (which was the case and should be communicated by the UI).

Some users asked how the collection of arguments was compiled. One user reported on an influence by having seen more pro than contra arguments, which indicated that both sides should be balanced. Another participant said that they were missing certain arguments, which

Figure 4.1: Comparison of one's own attitudes in ArgVote with the arguments of the parties (control group view, i.e. the user themself was not able to provide arguments).

means the possibility of adding own arguments, maybe even before the parties are asked for their attitudes, could make sense. Furthermore, transparency on the workings of the matching algorithm was asked for. In our experiment, it was important that participants did not know which algorithm is used; in a real-world application, however, this information should be provided with a thorough explanation why the algorithm is suitable and can be trusted. For this purpose, we conducted the experiment described in Section 3.3.

Further comments were made regarding the following issues:

- Arguments for neutral views should be provided.
- The UI for navigating back/forth the issues is hard or not obvious how to use.
- The importance of issues should be rated after having seen all of them to get a feeling which of them are more important.
- A tutorial explaining all features should be added. (Thus, the short introduction page in the questionnaire was not enough.)
- Conditional answers ("I agree, but only if . . . ") could be added. (For instance, there are different ways to reduce CO_2 emissions.)
- A mobile UI is missing.

All those points show that there are many open questions for future work, especially concerning the visual design and intuitiveness of the application as well as the compilation of arguments.

For the interpretation of the results of our experiments, most of the points mentioned above are not relevant, since they are systematic errors present in both groups.

4.2.4 Introducing a VAA Bot

We learned that having arguments in a VAA can improve the understanding of political positions and help with forming an opinion. But a major problem of asking for opinions on arguments in ArgVote is the increased time needed to complete the VAA: Participants which were exposed to arguments spent almost three times as much time in the application than the control group. This amount of time spent is okay if the users are interested in all the arguments and politics, but is unsuitable for people who want to get precise answers quickly.

Therefore, we contemplated another form of interaction which could reduce the perceived time spent within the VAA: We built a prototype for an interactive chat bot, which Brenneis and Mauve (2021a) presented in the following paper:

Markus Brenneis and Martin Mauve.
"ArgVote Bot: Introducing an Argumentative Voting Advice Bot"

Our contribution is as follows:

1. introduction of the idea of a VAA chat bot
2. presentation of a first implementation for such a bot

This VAA chat bot is our last application we present in this thesis. Similar to the original ArgVote web application, the chat bot raises interesting interdisciplinary questions, such as how it influences political interest and whether people trust an intelligent VAA.

Personal Contribution

Markus Brenneis, the author of this thesis, had the original idea of creating a VAA chat bot to mitigate the shortcomings of ArgVote revealed in the previous work presented in Subsection 4.2.1. He implemented the fulfillment server and designed the interaction flow in Dialogflow. The complete paper was written by Markus Brenneis. Martin Mauve, who had the original idea for an argument-based VAA, provided feedback on drafts of the paper.

Importance and Impact on This Thesis

The ArgVote chat bot is based on the general idea of an argument-based VAA presented in this section. It aims at addressing the limitations found in the experiment with the original ArgVote web application discussed in Subsections 4.2.1 and 4.2.3, in particular regarding interaction time and UI clarity.

ArgVote Bot: Introducing an Argumentative Voting Advice Bot

Markus Brenneis* and **Martin Mauve**
Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany
markus.brenneis@hhu.de

Abstract

Voting advice applications (VAAs) are used by a lot of people when they need help in deciding for which party to vote in the next election. But classical VAAs do not take into account *why* parties and voters have certain views. Therefore, we built an argument-based VAA, *ArgVote*, which considers the arguments used by parties and voters to calculate the party–voter similarity. As our first test with a classical web interface revealed that the presentation of arguments largely increased the interaction time, we propose a completely different kind of VAA: a VAA chat bot, which asks users about their attitudes and arguments concerning political positions and understands free text or spoken input. We use fastText embeddings to find arguments input by the users in our argumentation database.

1 Introduction

Voting advice applications (VAAs), e.g. *Vote Compass* or the German *Wahl-O-Mat*, are used by many people around the world [Van Camp *et al.*, 2014; Wagner and Ruusuvirta, 2012] to get support in deciding for whom to vote in an election. Those applications typically are websites asking for opinions concerning current political issues. Based on the user's and parties' answers, a voting advice is calculated. But classical VAAs do not consider *why* voters and parties have certain views. Consider, for example, the issue of nuclear power: It is a big difference whether a party is against nuclear power because they consider it unsafe, or because they think that the waste disposal is expensive.

To solve this shortcoming, we developed a new kind of argument-based VAA, *ArgVote* [Brenneis and Mauve, 2021], which also considers the reasons used to arrive at a position. For this purpose, our pseudometric [Brenneis *et al.*, 2020] for calculating the similarity of attitudes in argumentations is used. Our first experiment with a web interface revealed that presenting arguments significantly increases the understanding of political topics and different opinions. But we also saw that the inclusion of arguments increased the time needed to interact with the VAA by a factor of 3, and that users considered the resulting interface crowded. Therefore, we created a new and completely different approach: a VAA chat bot, which should regularly ask users about their views before an election and can understand the users' arguments which are entered as free text.

The idea behind our chat bot is that it can collect the information needed for a voter–party comparison in a natural language conversion over a certain period of time. The advantage compared to using a web app at a dedicated point of time is that it does not involve a user interface with many elements and can reduce the time needed for gathering the pieces of information.

Our chat bot idea is inspired by previous work. *jebediah* [Meter *et al.*, 2018] is a chat bot based on Google's Dialogflow[1] platform, which enables users to communicate with the Dialog-based argumentation system (D-BAS) [Krauthoff *et al.*, 2017]; they can argue about different positions from within existing platforms like Facebook and Telegram and do not have to use D-BAS' web interface. Another, persuasive chat bot was created by Chalaguine and Hunter [2020]. Their bot attempts to understand user concerns and tries to find matching arguments in its argumentation database using Cosine similarity. One other decision helping tool is the online platform Cannabis Convo[2]. Users can find out about their stance regarding the cannabis reform in New Zealand by giving opinions on different aspects of the process and assessing their relevance and see the consequences of their opinion.

In the next section, we will explain which features our chat bot has. Afterwards, we have a look at the implementation of the bot. Last of all, we summarize our contribution and look at future work.

2 Features of Our Chat Bot

In our ArgVote bot, voters can text (or talk) with a bot over some period (e.g. one topic every morning, starting four weeks before an election) and are asked by the bot about their opinions and arguments regarding a number of political positions. This form of communication might feel like a more natural recommender than a dedicated platform, and users are not confronted with many issues and arguments at once.

*Contact Author

[1] https://dialogflow.cloud.google.com/
[2] https://www.cannabisconvo.co.nz/

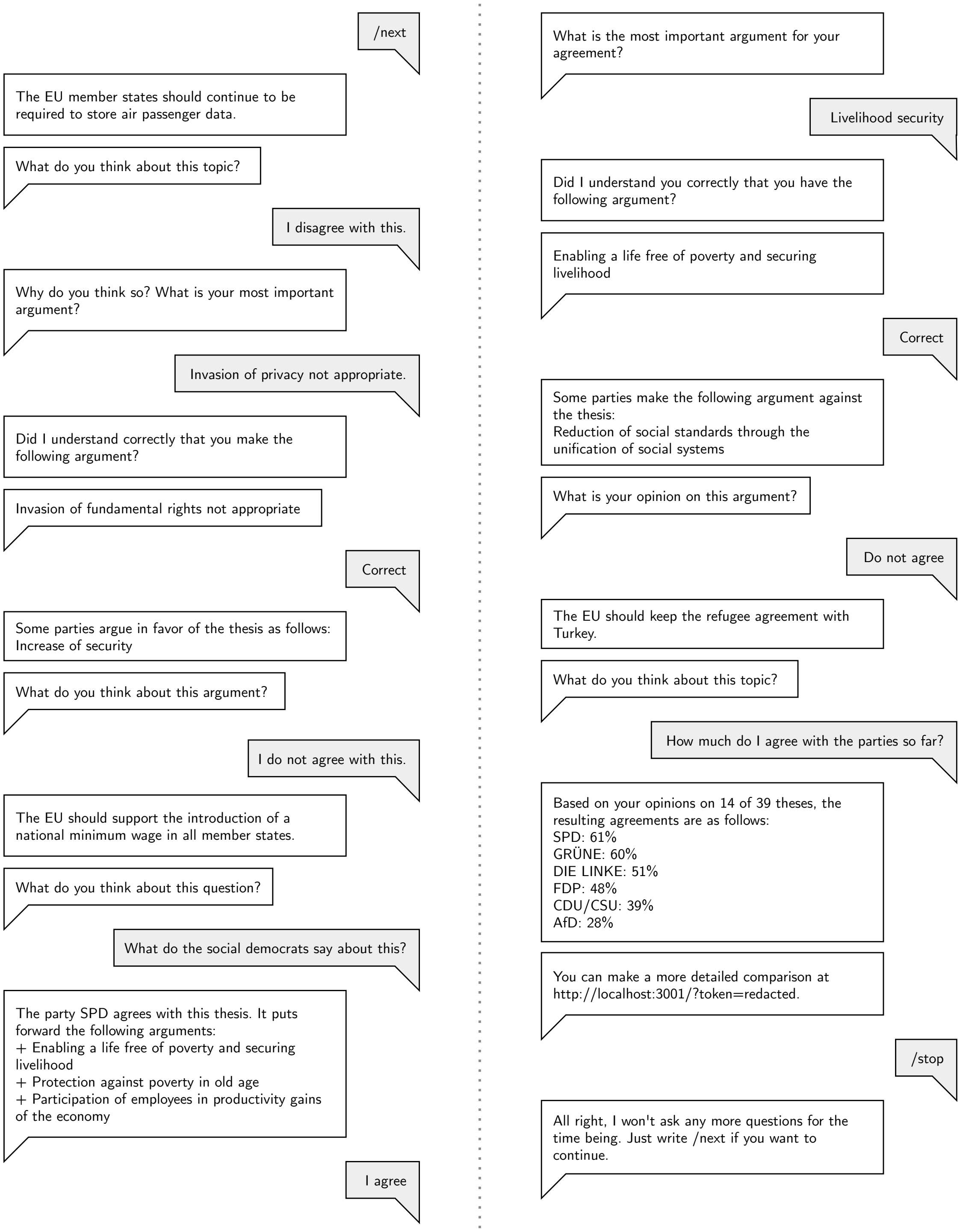


Figure 1: A conversation between our ArgVote bot (left) and a user (right). The original German text was translated to English as literally as possible to demonstrate the performance of the argument matching.

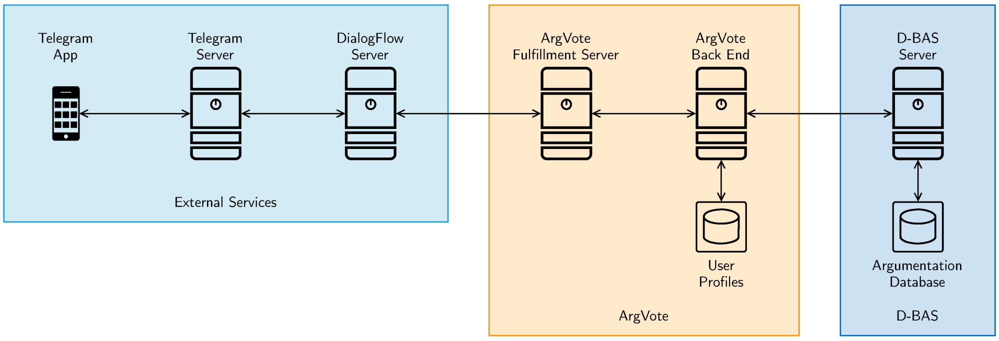


Figure 2: Overview of the communicating services for ArgVote's Telegram integration. The ArgVote Fulfillment Server is the main contribution of this work and is responsible for generating appropriate answers to the users' input.

Our bot has the following features:

- asking users for their opinions concerning political positions
- displaying a party's attitude (incl. arguments) on a position
- showing arguments for/against a position
- asking users for their most important argument for their opinion
- presenting a rebut to a user's argument and asking for the user's opinion on that argument
- displaying the matching with the parties
- changing a previously entered opinion on a position

An example dialog demonstrating most of these features is depicted in Figure 1. When designing the bot's responses, we made sure to communicate what the bot understood. This explicitness helps to prevent that wrong information is saved in case of a misunderstanding.

3 Implementation Details

Our software is based on the DialogFlow platform. Determining a user's intent to an input is done by DialogFlow using training phrases provided by us, e.g. "I do not agree with this." or "I disagree." For (most) matched intents (those which require information from ArgVote), a request is sent from DialogFlow to our fulfillment server to handle the request. Possibly recognized, intent-dependent parameters, like opinions or party names, are included in the request. The fulfillment server then queries necessary information from ArgVote's back end, and replies with the answer which should be presented to the user, including quick-reply buttons (e.g. "I agree") for supporting platforms.

Users' attitudes are saved in ArgVote, such that switching to its original, full web front end is always possible for the user. The arguments are stored in an instance of our D-BAS system to enable sharing arguments with other front ends based on D-BAS.

The fulfillment server itself is stateless, which allows a simple design. Our intent design makes sure that DialogFlow keeps track of the context of a conversation (e.g. the id of the position currently talked about) and includes the context in follow-up requests.

For matching the arguments entered by the user with the arguments in ArgVote's data base, the Cosine similarity of the averaged fastText embedding [Grave *et al.*, 2018] is calculated by our fulfillment server. We use pre-trained word vectors based on Common Crawl[3]. This approach is similar to the argument matching algorithm by Chalaguine and Hunter [2020], which was successfully tested in their chat application.

DialogFlow allows integrating our bot with different chat or voice front ends. Figure 2 summarizes which services communicate with each other in case of the integration in the Telegram messenger.

4 Summary & Future Work

We developed a VAA chat bot, which asks users for their political views and the arguments for their stances. The free-text arguments entered by the users are matched with the bot's argumentation database using word embeddings. Based on the user information, a voting advice is calculated. Moreover, the bot can answer questions about the views of political parties.

We have not yet tested the perception of our bot "in the wild." In particular, we do not know how good the text matching of arguments entered by users works. We anticipate that this new form of interaction could increase the awareness and knowledge of political issues and might improve the turnout at an election. On the other hand, some people might be scared due to a negative attitude towards chat bots or artificial intelligence and could get the feeling of being manipulated. Those effects have to be studied thoroughly.

References

[Brenneis and Mauve, 2021] Markus Brenneis and Martin Mauve. ArgVote: Which Party Argues Like Me? Exploring an Argument-Based Voting Advice Application. In I. Czarnowski, R. J. Howlett, and L. C. Jain, editors, *Intelligent Decision Technologies 2021 – Proceedings of the 13th KES International Conference on Intelligent Decision*

[3]https://fasttext.cc/docs/en/crawl-vectors.html

Technologies (KES-IDT-21), Singapore, 2021. Springer. To appear.

[Brenneis *et al.*, 2020] Markus Brenneis, Maike Behrendt, Stefan Harmeling, and Martin Mauve. How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs. In Sarah A. Gaggl, Matthias Thimm, and Mauro Vallati, editors, *Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020)*, number 2672 in CEUR Workshop Proceedings, pages 2–13, Aachen, September 2020.

[Chalaguine and Hunter, 2020] Lisa A. Chalaguine and Anthony Hunter. A persuasive chatbot using a crowd-sourced argument graph and concerns. *Frontiers in Artificial Intelligence and Applications*, 326:9–20, 2020.

[Grave *et al.*, 2018] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[Krauthoff *et al.*, 2017] Tobias Krauthoff, Christian Meter, and Martin Mauve. Dialog-Based Online Argumentation: Findings from a Field Experiment. In *Proceedings of the 1st Workshop on Advances in Argumentation in Artificial Intelligence*, pages 85–99, November 2017.

[Meter *et al.*, 2018] Christian Meter, Björn Ebbinghaus, and Martin Mauve. Jebediah – Arguing With a Social Bot. In *Frontiers in Artificial Intelligence and Applications, Computational Models of Argument*, volume 305, pages 467–468. IOS Press, 9 2018.

[Van Camp *et al.*, 2014] Kirsten Van Camp, Jonas Lefevere, and Stefaan Walgrave. The content and formulation of statements in voting advice applications. *Matching Voters With Parties and Candidates. Voting Advice Applications in Comparative Perspective*, pages 11–32, 2014.

[Wagner and Ruusuvirta, 2012] Markus Wagner and Outi Ruusuvirta. Matching voters to parties: Voting advice applications and models of party choice. *Acta politica*, 47(4):400–422, 2012.

Chapter 5

Conclusion

In this chapter, we summarize the main results and contributions of this thesis and look at work for the future. We dealt with the question of how to compare the attitudes of different participants in an argumentation and what applications there are for such a comparison in the context of online argumentations. A theoretical model to capture attitudes in argumentations was developed, and we proposed a pseudometric to calculate the distance between attitudes. Moreover, we found out how much different distance functions agree with human assessments of argumentation similarity. Afterwards, two application scenarios in which our pseudometric has been used were studied: First, we looked at a recommender engine for arguments, and secondly, at an argument-based VAA.

During the work for this thesis, five full papers (Brenneis et al., 2021, 2020; Brenneis and Mauve, 2020b, 2021b,c) and one paper for a demo session (Brenneis and Mauve, 2020a) were accepted for publication; in addition, one technical report (Brenneis and Mauve, 2021a) was written. We developed the new applications *deliberate* (cf. Section 4.1) and ArgVote (cf. Section 4.2). Furthermore, three datasets, which may be used by the research community, were released:

- Human similarity assessments for several argumentation scenarios[1]
- Argumentation attitude dataset on two different positions from different points of time[2]
- Argumentation tree and user profiles from our argument-based VAA[3]

5.1 Main Results

This sections sums up the key results of our work. To begin with, we consider the theoretical work about developing a metric for comparing attitudes in argumentations. Then, we wrap up what we learned from the use cases of our pseudometric in *deliberate*, the UPEKI project, and our experiment with ArgVote.

[1] `https://github.com/hhucn/argumentation-similarity-survey-results`, cf. Section 3.3 and Brenneis and Mauve (2020b)

[2] `https://github.com/hhucn/argumentation-attitude-dataset`, cf. Subsection 4.1.5 and Brenneis et al. (2021)

[3] `https://github.com/hhucn/argvote-dataset`, cf. Section 4.2 and Brenneis and Mauve (2021b)

5.1.1 Theory

In Chapter 3, we took an extensive look at how the attitudes of people who exchange arguments can be compared to get a similarity ranking. We first developed the model of weighted argumentation graphs to capture all relevant pieces of information which arise in real-world discussions: the weight of arguments and positions, as well as the rating of statements.

Then we presented our pseudometric, which calculates the distance between two weighted argumentation graphs. The main property of this distance function is that it decreases the contribution of statements which are further away from the positions. To assure that it yields sensible results, we proposed and proved the fulfillment of nine desiderata.

Thereafter, we asked how intuitive our desiderata, and thus our pseudometric, are and presented an empirical study to find out how average humans perform comparisons of argumentations. We had a list of 33 hypotheses regarding basic properties of argumentations, influence of weights, influence of missing information, and trade-off situations. Our study yielded some surprising results, i.a. we discovered that people often followed a triangle model for pro–neutral–contra.

Last of all, we compared different distance functions with our human baseline results. We included seven different functions in our comparison and pointed out why certain functions performed better or worse. We saw that, depending on the application context and its peculiarities, our pseudometric or simpler distance functions like the p-metric performed best.

5.1.2 Applications

After that theoretic view, we considered possible applications for the pseudometric in Chapter 4. We presented the web application *deliberate* with an argument recommender system. This system uses a nearest-neighbor algorithm based on our pseudometric to pre-filter arguments. The goal of the filtering process is making larger argumentations clearer and more manageable by only displaying arguments relevant for a user.

The application was used in two studies in the UPEKI project, in which the influence of different filtering strategies on the formation of opinion is researched. Based on the empirical studies of the research group, we created a dataset for evaluating argument recommender systems. We compared a majority baseline algorithm with deliberate's algorithm and found out that the latter performed better, but there is still room for improvement.

Finally, we introduced our new, argument-based VAA ArgVote, which can display arguments next to political theses. It considers the opinions on arguments in its matching algorithm and uses our pseudometric to calculate the voter–party similarity. In an empirical study, we checked the general acceptance of our VAA. We discovered that including arguments improved the understanding of political topics and different opinions, and users liked using the application. But our matching algorithm could not be considered better than a classical VAA matching algorithm. Moreover, we drafted the idea of a chat bot for ArgVote which aims at reducing the perceived time users need to interact with the theses and their arguments.

5.2 Future Work

We provide several new methods and applications for comparing the attitudes of different participants in argumentations. But as we already pointed out, there are various open questions which can be dealt with in the future.

We suggested a new pseudometric for calculating the dissimilarity of weighted argumentation graphs. Based on our human baseline, one can try to find other, possibly "real" metrics, which fully match human intuition, and adapt them to other models of argumentation, e.g. GAFs. Furthermore, the human baseline results can be extended with surveys including more participants from different countries and age groups. This way, we can find out whether there are any differences in the assessments of argumentation similarity between those groups, since we currently only have data from the USA. Also, we should get more significant and probably clearer results for argumentation scenarios for which we do not know a definite answer, yet, especially with regard to deeper and bigger scenarios. One should also investigate how the questions are presented to untrained humans to not ask too much of them, for instance regarding deeper argumentations or undercuts. Moreover, further results and scenarios could be used to tune hyperparameters of distance functions, e.g. how much deeper arguments contribute to the overall difference.

For all distance functions which can be defined, it is essential to remember that they must prove useful in an application context. Besides the two applications we have looked at, one can explore more use cases, for example automatically finding a consensus, clustering of voters, or computational persuasion (Hunter, 2018). Another factor to keep in mind is that a software must be able to collect the information necessary to perform a comparison from the users, but the UI may not become too difficult to use.

In the context of an argument recommender engine, we focused on presenting convincing arguments. Other goals could be investigated, for instance seeing arguments for opinions different from one's own or providing support for defending one's own point of view. Moreover, other algorithms can now be benchmarked against our new dataset. New algorithms might include linguistic features, different distance functions, or another machine learning model. What is more, an open question is how easy "evil" users can manipulate such a system.

Last but not least, our idea of an argument-based VAA can be studied more thoroughly. As we now provide a dataset with user profiles, the results of different distance measures can be compared with the subjective party sympathy of the users. Gathering a bigger dataset would enable studying more sophisticated matching algorithms, as used in Social VAAs. A streamlined UI, which should also be mobile-friendly, could be tested in a comparative study. One also has to find a way to reduce the time needed to interact with the VAA, possibly through a completely new interaction model. For example, the bot solution we proposed has not yet been tested with real users. It could also be considered that the bot prioritizes questions which better help to discern the most matching parties. When creating a bot, questions about neutrality and fears about artificial intelligence must be taken into account.

Another major question for a real-life application of the VAA is where the arguments come from, how they are phrased, and how to make sure pro and contra arguments are balanced. The inclusion of conditional statements is an issue not only interesting for ArgVote, but also for classical VAAs. The same is true of investigating how easily parties can manipulate rankings by providing dishonest answers or not answering some questions. Lastly, we only tested ArgVote with a small, young, highly educated sample of the German population, thus, we do not know how displaying arguments is perceived by e.g. older people and people from other countries. As we can see, the whole topic of argument-based VAAs has great potential for interdisciplinary research.

5.3 Closing Thoughts

Within our argumentation research group, we have the vision of treating arguments as a valuable resource. It should be easy to reuse arguments in different contexts to prevent tiring repetitions as well as missing and incomplete arguments. Consider how often you might have explained in different conversations and on different online platforms why vegans are not destroying the rain forest and how often you had to dig up the same references again and again.

If all arguments ever made were simply available in a web of arguments, those repetitions would no longer be necessary. EDEN goes one step into providing such an infrastructure. *discuss* is an example of gathering the content, i.e. the arguments, user-friendly.

Taking the idea even further, not only the arguments, but also one's attitudes are valuable. You could carry your personal argumentation graph on a (metaphorical) pen drive and plug it in when you want to share it with applications like *deliberate* and ArgVote to get a personalized experience or a party matching result without having to enter your opinions again.

If public entities like political parties made their attitudes publicly available, this would, in theory, allow everyone at every time to compare one's own attitude with the attitude of parties and other public institutions. If users also shared parts of their personal argumentation graphs, public entities would know what the population thinks and which policy would maximize the satisfaction of all citizens.[4]

[4] We are well aware that this vision suffers from some practical problems ranging from privacy over manipulation to feasibility, as already partly discussed in this work.

Glossary

Amazon Mechanical Turk (MTurk) Amazon Mechanical Turk is a crowdsourcing platform where remote workers can be hired to perform small tasks. Workers can be selected e.g. based on region and language skills. 44

Application Programming Interface (API) An API defines how multiple pieces of software can interact with each other. 90

Argument An argument is the connection of two statements (or, in case of an undercut, a statement and an argument), which are called premise and conclusion, e.g. "More nuclear power plants should be built because they do not cause any emissions." In everday language, the premise statement alone is sometimes referred to as argument, but the argumentative context is only established when there is some connection to a conclusion. 3, 5, 6, 10–14, 16, 18, 19, 33, 35, 37, 39, 63, 65, 87–89, 94, 108, 122, 132

Argumentation Graph An argumentation graph is a data structure which captures the relations of statements, and thus arguments, in an argumentation. In this work, nodes are statements, edges are arguments. 6

ArgVote ArgVote is a prototype for an argument-based VAA developed by Brenneis and Mauve (2021b) which does not only ask users for their opinions on political theses, but also for their arguments. The opinions on arguments are considered when calculating the party–user similarity. 4, 89, 108, 109, 120–126, 131–134, 149

Attack A premise attacks its conclusion if it argues against the conclusion, as in "Nuclear power should be banned because we do not know how nuclear waste can be stored safely in the long term." To be precise, there are different kinds of attacks, namely rebut, undermine, and undercut, which are discerned based on the relation to a previously mentioned argument. 2, 5, 6, 8, 12, 13, 39, 96

Attitude The attitude is the sum of a user's personal ratings and weights for an argumentation graph. 1–4, 9–11, 13, 14, 18, 32, 33, 35, 38, 39, 41, 43, 64, 68, 89, 94, 99, 108, 120, 122, 125, 131–134

Bipolar Weighted Argumentation Frameworks (BWAF) In a Bipolar Weighted Argumentation Framework, as introduced by Pazienza et al. (2017a), there are non-zero weights for relations, whose sign determines whether the relation is attacking or defending. 13

Claim *see* conclusion 5

Collaborative Filtering Collaborative filtering is a method used by recommender systems that tries to predict user preferences by collecting information about many users and looking at the preferences of similar users. 2, 4, 89, 90, 94, 121

Conclusion A conclusion (or claim) is a statement (or argument, see undercut) which is supported or attacked by another statement (which is called premise) to form an argument; for example, the argument "Nuclear power should be banned because we do not know how nuclear waste can be stored safely in the long term." has the conclusion "Nuclear power should be banned." 5–7, 11, 12, 16, 34, 37, 38, 87

decide The web application *decide* has been developed by Ebbinghaus and Mauve (2020) and supports collaborative decisions after an exchange of arguments. 8, 13, 136

Defend A premise defends (or supports) its conclusion if it argues in favor of the conclusion, as in "Nuclear power should not be banned because it is a sustainable energy source." 2, 5

deliberate Brenneis and Mauve (2020a) developed *deliberate*, a full-stack web application for exchanging arguments on a position, which includes a recommender system for arguments. 4, 19, 89–91, 94–99, 109, 120, 121, 131, 132, 134, 151

Dialog-Based Argumentation System (D-BAS) D-BAS is a full-stack web application developed by Krauthoff et al. (2018) which simulates a time-shifted, chat-like argumentation between different people. 1, 7, 8, 12, 89, 90, 96, 149, 151

discuss Meter et al. (2017) developed *discuss* to embed dialog-based argumentation into websites. 8, 95, 96, 134, 151

Distance Function In this thesis, a distance function is any function which calculates some kind of distance between two points. In a stricter meaning, *distance function* is a synonym for *metric*. 2–4, 9, 14, 18, 32, 44, 64–69, 87, 89, 122, 123, 131–133, 151

Evidence *see* premise 5

Extensible Discussion Entity Network (EDEN) EDEN was developed by Meter et al. (2018b) to provide a distributed argumentation graph which can be used by different front ends like *discuss*, *deliberate* or D-BAS. 8, 134

Family-Wise Error Rate (FWER) The family-wise error rate is the probability of making at least one type I error when testing multiple hypotheses at once based on the same observed set of values. 61

Generalized Argumentation Framework (GAF) GAFs were introduced by Ferilli (2020) and are based on BWAFs. They comprise bipolar argument relations, weights for nodes and edges, and additional external information. 13, 14, 133

Intersection–Union Test (IUT) An intersection–union test is a test method used in hypothesis testing which can be applied when the null hypothesis is a union (disjunction) of sub-hypotheses. 61, 63

Issue In the context of argumentation graphs following the IBIS model, the issue (commonly abbreviated as I) is the root of such a graph and represents the overal topic of an argumentation, e.g. “combating climate change.” 6, 11, 94

Issue-Based Information System (IBIS) IBIS is an argumentation system developed by Kunz and Rittel (1970) and built around positions related to a common issue. 6, 10, 11

Jebediah Jebediah is a chat bot based on D-BAS developed by Meter et al. (2018a). 8

Metric A metric is a distance function, satifsying identity of indiscernibles, symmetry, and the triangle inequality. 2, 3, 12, 14, 18, 19, 32, 36, 37, 41, 43, 64, 122, 131, 133

Personal well-being The personal well-being, also referred to as I, is our alternative interpretation of the issue in the IBIS model, which allows interpreting positions as pseudo-arguments which can have weights. For instance, one could say “My personal well-being is improved, because nuclear power is banned.” 11

Position Statements which have no conclusions are called positions, are typically actionable items like “Nuclear power should be banned,” and are drawn as pseudo-premise of the issue I. 3, 6, 7, 10–12, 14, 15, 32–43, 64, 67, 87, 88, 94–97, 122, 131, 132

Premise A premise (or evidence) is a statement which supports or attacks another statement or argument (which is called conclusion) to form an argument; for example, the argument “Nuclear power should be banned because we do not know how nuclear waste can be stored safely in the long term.” has the premise “We do not know how nuclear waste can be stored safely in the long term.” 5–7, 13, 16, 37, 39, 94, 122

Proposition *see* statement 5

Pseudometric A pseudometric d is a metric, but points need not to be distinguishable, i.e. $x \neq y \not\Rightarrow d(x, y) \neq 0$. For example, Brenneis et al. (2020) presented a pseudometric for weighted argumentation graphs. 3, 4, 9, 12, 14, 15, 18, 19, 32, 33, 39, 44, 60, 64, 65, 68, 69, 87, 88, 90, 91, 94, 99, 109, 120, 122, 123, 131–133

Rank-Biased Overlap (RBO) The RBO developed by Webber et al. (2010) is a measure for the similarity of ranked lists, which considers differences at the top-positions more severe than differences in lower positions. 123

Rating In a weighted argumentation graph, a rating is a numeric value from -0.5 to 0.5 a person can assign to a statement to indicate how sure they are that the statement is true. 10–16, 18, 63, 64, 67, 94, 95, 122, 124, 132

Rebut A rebut is an attack on an attacking/supporting argument by bringing forward an argument which supports/attacks its conclusion, e.g. the argument “Nuclear power should be banned because nuclear power plants are insecure.” can be rebutted with “Nuclear power should not be banned because it is sustainable.” 5

Recommender System A recomender system is a program which predicts a value such as a rating a user would assign to an item, e.g. using collaborative filtering. 3, 4, 9, 19, 32, 89–91, 94, 95, 99, 132

Statement A statement (or proposition) is the atomic building block of arguments, i.e. either a premise or a conclusion. 3, 5–7, 10–19, 33, 39, 63, 65, 68, 87, 132

Support *see* defend 5, 6, 96

Undercut An undercut is an attacking argument which has an argument as conclusion, e.g. the argument "Eating meat is okay because lions eat meat." can be undercut with the argument "This is a naturalistic fallacy."; the attack does not question the validity of the premise and the conclusion, but the argumentative relation. 5, 6, 8, 12, 13, 39, 121, 133

Undermine An undermine is an attack on an argument's premise, e.g. the argument "Nuclear power should be banned because nuclear power plants are insecure." can be undermined with "Nuclear power plants are secure because there are safety precautions." 5, 12

Unterstützung politischer Entscheidungen durch Künstliche Intelligenz (UPEKI) UPEKI is the political use case of the Manchot research group *Decision-making with the help of Artificial Intelligence*, which focuses on AI support for policy decisions and deals with the development of tools and impact analysis. 4, 89–91, 94–99, 124, 131, 132

User Interface (UI) In the context of this work, the UI comprises the visual parts of a software an end user interacts with. 3, 4, 7, 8, 12, 96, 97, 109, 122, 124–126, 133

Voting Advice Application (VAA) A VAA is an application which can be used by voters to find out which political parties match their own preferences most. Popular examples are the German Wahl-O-Mat and Election Compass USA. 4, 14, 19, 32, 44, 65, 67–69, 87–89, 96, 97, 108, 109, 120–124, 126, 131–134

Weight In a weighted argumentation graph, a weight is a numeric value from 0 to 1 a person can assign to an argument to indicate how important they consider this argument in comparison to other arguments with the same conclusion. 11–16, 18, 32, 35, 38, 39, 63–65, 122, 132

Weighted Argumentation Graph A weighted argumentation graph, as defined by Brenneis et al. (2020), is an argumentation graph which additionally captures a user's weights of arguments and ratings of statements. 3, 4, 9–12, 14, 19, 32, 65, 120, 122, 132, 133

Bibliography

Amgoud, Leila, Jonathan Ben-Naim, Dragan Doder, and Srdjan Vesic (2017). "Acceptability Semantics for Weighted Argumentation Frameworks". In: *IJCAI*. Vol. 2017, pp. 56–62 (cit. on pp. 6, 13).

Amgoud, Leila and Victor David (2020). "An Adjustment Function for Dealing with Similarities". In: *Computational Models of Argument*. Vol. 326. IOS Press, pp. 79–90 (cit. on p. 32).

Arrow, Kenneth J (1950). "A difficulty in the concept of social welfare". In: *Journal of political economy* 58.4, pp. 328–346 (cit. on p. 32).

Bench-Capon, Trevor and Sanjay Modgil (2009). "Case law in extended argumentation frameworks". In: *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pp. 118–127 (cit. on p. 6).

Berger, Roger L. and Jason C. Hsu (1996). "Bioequivalence Trials, Intersection–Union Tests and Equivalence Confidence Sets". In: *Statistical Science* 11.4, pp. 283–319 (cit. on p. 61).

Bhavsar, Virendrakumar C., Harold Boley, and Lu Yang (2004). "A Weighted-Tree Similarity Algorithm for Multi-Agent Systems in E-Business Environments". In: *Computational Intelligence* 20.4, pp. 584–602 (cit. on pp. 64, 66, 122).

Bille, Philip (2005). "A survey on tree edit distance and related problems". In: *Theoretical computer science* 337.1-3, pp. 217–239 (cit. on p. 14).

Booth, Wayne C., Gregory G. Colomb, and Joseph M. Williams (2008). *The craft of research*. 3rd ed. University of Chicago press (cit. on p. 5).

Brenneis, Markus, Maike Behrendt, and Stefan Harmeling (July 2021). "How Will I Argue? A Dataset for Evaluating Recommender Systems for Argumentations". In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Singapore and Online: Association for Computational Linguistics, pp. 360–367 (cit. on pp. 98, 131, 147).

Brenneis, Markus, Maike Behrendt, Stefan Harmeling, and Martin Mauve (Sept. 2020). "How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs". In: *Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020)* (Sept. 8, 2020). Ed. by Sarah A. Gaggl, Matthias Thimm, and Mauro Vallati. CEUR Workshop Proceedings 2672. Aachen, pp. 2–13 (cit. on pp. 10, 14, 65, 131, 137, 138, 147).

Brenneis, Markus and Martin Mauve (Sept. 2020a). “deliberate – Online Argumentation with Collaborative Filtering”. In: *Computational Models of Argument*. Ed. by Henry Prakken, Stefano Bistarelli, Santini Francesco, and Carlo Taticchi. Vol. 326. IOS Press, pp. 453–454 (cit. on pp. 89, 131, 136, 147).

Brenneis, Markus and Martin Mauve (Nov. 2020b). “Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations”. In: *Proceedings of the Workshop on Advances In Argumentation In Artificial Intelligence 2020* (Nov. 25, 2020). Ed. by Bettina Fazzinga, Filippo Furfaro, and Francesco Parisi. CEUR Workshop Proceedings 2777. Aachen, pp. 1–15 (cit. on pp. 33, 34, 64, 131, 147).

Brenneis, Markus and Martin Mauve (2021a). *ArgVote Bot: Introducing an Argumentative Voting Advice Bot*. Technical Report: Heinrich Heine University Düsseldorf (cit. on pp. 126, 131, 147).

Brenneis, Markus and Martin Mauve (July 2021b). “ArgVote: Which Party Argues Like Me? Exploring an Argument-Based Voting Advice Application”. In: *Intelligent Decision Technologies*. Ed. by I. Czarnowski, R. J. Howlett, and L. C. Jain. Springer Singapore, pp. 3–13 (cit. on pp. 89, 108, 123, 131, 135, 147).

Brenneis, Markus and Martin Mauve (July 2021c). “How Intuitive Is It? Comparing Metrics for Attitudes in Argumentation with a Human Baseline”. In: *Artificial Intelligence in HCI*. Ed. by Helmut Degen and Stavroula Ntoa. Springer International Publishing, pp. 125–138 (cit. on pp. 64, 68, 131, 147).

BVerfG (2002). *Judgment of the First Senate of 17 July 2002 - 1 BvF 1/01 -, paras. 1-147*. URL: `http://www.bverfg.de/e/fs20020717_1bvf000101en.html` (cit. on p. 121).

Cayrol, Claudette and Marie-Christine Lagasquie-Schiex (2005). “On the acceptability of arguments in bipolar argumentation frameworks”. In: *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer, pp. 378–389 (cit. on p. 6).

Chalaguine, Lisa A. and Anthony Hunter (2020). “A Persuasive Chatbot Using a Crowd-Sourced Argument Graph and Concerns”. In: *Frontiers in Artificial Intelligence and Applications* 326, pp. 9–20 (cit. on pp. 1, 8).

Ciaccia, Paolo, Marco Patella, and Pavel Zezula (1997). “M-Tree: An Efficient Access Method for Similarity Search in Metric Spaces”. In: *Proceedings of the 23rd International Conference on Very Large Data Bases*. VLDB ’97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 426–435 (cit. on p. 32).

Collenette, Joe, Katie Atkinson, and Trevor Bench-Capon (2020). “An Explainable Approach to Deducing Outcomes in European Court of Human Rights Cases Using ADFs”. In: *Frontiers in Artificial Intelligence and Applications* 326, pp. 21–32 (cit. on p. 6).

Correnz, Fabian (Dec. 2020). “Entwicklung eines Chat-Bots für das Argumentationssystem D-BAS”. Bachelor’s Thesis. Department of Computer Science, Heinrich-Heine-University Düsseldorf (cit. on p. 8).

Damer, T. Edward (Feb. 2008). *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*. 6th ed. (cit. on p. 13).

Daxenberger, Johannes and Iryna Gurevych (2020). “Arguments as Social Good: Good Arguments in Times of Crisis”. Ubiquitous Knowledge Processing (UKP) Lab, Technische Universität Darmstadt (cit. on p. 1).

Dung, Phan Minh (1995). “On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games”. In: *Artificial Intelligence* 77.2, pp. 321–357 (cit. on pp. 6, 13).

Dunne, Paul E., Anthony Hunter, Peter McBurney, Simon Parsons, and Michael Wooldridge (2011). “Weighted argument systems: Basic definitions, algorithms, and complexity results”. In: *Artificial Intelligence* 175.2, pp. 457–486 (cit. on p. 13).

Ebbinghaus, Björn (May 2019). “Decision Making with Argumentation Graphs”. Master’s Thesis. Department of Computer Science, Heinrich-Heine-University Düsseldorf (cit. on pp. 1, 12).

Ebbinghaus, Björn and Martin Mauve (Sept. 2020). “decide: Supporting Participatory Budgeting with Online Argumentation”. In: *Computational Models of Argument. Proceedings of COMMA 2020*. Ed. by Henry Prakken, Stefano Bistarelli, Santini Francesco, and Carlo Taticchi. Vol. 326. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 463–464 (cit. on pp. 8, 13, 136).

Elahi, Mehdi, Francesco Ricci, and Neil Rubens (2016). “A survey of active learning in collaborative filtering recommender systems”. In: *Computer Science Review* 20, pp. 29–50 (cit. on p. 94).

Ferilli, Stefano (Nov. 2020). “Towards a General Model for Abstract Argumentation Frameworks”. In: *Proceedings of the Workshop on Advances In Argumentation In Artificial Intelligence 2020*. Ed. by Bettina Fazzinga, Filippo Furfaro, and Francesco Parisi. CEUR Workshop Proceedings 2777. Aachen, pp. 60–74 (cit. on pp. 13, 136).

Franklin, Charles H. (2007). *The ‘Margin of Error’ for Differences in Polls*. URL: `https://abcnews.go.com/images/PollingUnit/MOEFranklin.pdf` (cit. on p. 60).

Freeman, James B. (2011). *Argument Structure: Representation and Theory*. Vol. 18. Springer Science & Business Media (cit. on p. 5).

Gordon, Thomas F. (2013). “Introducing the Carneades web application”. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pp. 243–244 (cit. on p. 7).

Gordon, Thomas F. and Douglas Walton (2016). "Formalizing Balancing Arguments". In: *Frontiers in Artificial Intelligence and Applications* 287, pp. 327–338 (cit. on pp. 6, 13).

Hadjinikolis, Christos, Yiannis Siantos, Sanjay Modgil, Elizabeth Black, and Peter McBurney (2013). "Opponent Modelling in Persuasion Dialogues". In: *Twenty-Third International Joint Conference on Artificial Intelligence* (cit. on p. 94).

Holm, Sture (1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian journal of statistics*, pp. 65–70 (cit. on p. 61).

Hunter, Anthony (2013). "A probabilistic approach to modelling uncertain logical arguments". In: *International Journal of Approximate Reasoning* 54.1, pp. 47–81 (cit. on pp. 6, 13).

Hunter, Anthony (2018). "Towards a framework for computational persuasion with applications in behaviour change". In: *Argument & Computation* 9.1, pp. 15–40 (cit. on p. 133).

Hunter, Anthony, Sylwia Polberg, and Matthias Thimm (2020). "Epistemic graphs for representing and reasoning with positive and negative influences of arguments". In: *Artificial Intelligence* 281, p. 103236 (cit. on pp. 6, 13).

Ittermann, Iwan and Brian Plüss (2020). "PEOPLES: From private responses to messages to depolarisation nudges in two-party adversarial online talk". In: *8th International Conference on Computational Models of Argument, COMMA 2020*. IOS Press, pp. 467–468 (cit. on p. 8).

Katakis, Ioannis, Nicolas Tsapatsoulis, Fernando Mendez, Vasiliki Triga, and Constantinos Djouvas (2013). "Social Voting Advice Applications - Definitions, Challenges, Datasets and Evaluation". In: *IEEE Transactions on Cybernetics* 44.7, pp. 1039–1052 (cit. on p. 109).

Kelm, Ole, Maike Behrendt, Markus Brenneis, Katharina Gerl, Stefan Harmeling, Stefan Marschall, Florian Meißner, Tim Neumann, Gerhard Vowe, Carina Weinmann, and Marc Ziegele (2021). "How algorithmically curated online environments influence users' political polarization". Publication planned (cit. on pp. 95, 98).

Krauthoff, Tobias, Christian Meter, Gregor Betz, Michael Baurmann, and Martin Mauve (Sept. 2018). "D-BAS – A Dialog-Based Online Argumentation System". In: *Computational Models of Argument*. Warsaw, pp. 325–336 (cit. on p. 136).

Krauthoff, Tobias, Christian Meter, and Martin Mauve (Nov. 2017). "Dialog-Based Online Argumentation: Findings from a Field Experiment". In: *Proceedings of the 1st Workshop on Advances in Argumentation in Artificial Intelligence*. Bari, pp. 85–99 (cit. on pp. 1, 7, 12).

Kunz, Werner and Horst W. J. Rittel (1970). *Issues as elements of information systems*. Vol. 131. Citeseer (cit. on pp. 6, 10, 137).

Kunze, Matthias, Matthias Weidlich, and Mathias Weske (2011). "Behavioral Similarity – A Proper Metric". In: *International Conference on Business Process Management*. Springer, pp. 166–181 (cit. on p. 65).

Lawrence, John and Chris Reed (Jan. 2020). “Argument Mining: A Survey”. In: *Comput. Linguist.* 45.4, pp. 765–818 (cit. on pp. 1, 7).

Liebeck, Matthias, Katharina Esau, and Stefan Conrad (2016). “What to do with an airport? Mining arguments in the German Online Participation Project Tempelhofer Feld”. In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pp. 144–153 (cit. on pp. 1, 7).

Martínez, Diego C., Alejandro J. García, and Guillermo R. Simari (2008). “An Abstract Argumentation Framework with Varied-Strength Attacks”. In: *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR'08)*, pp. 135–144 (cit. on p. 13).

Meter, Christian, Björn Ebbinghaus, and Martin Mauve (Sept. 2018a). “Jebediah – Arguing With a Social Bot”. In: *Frontiers in Artificial Intelligence and Applications, Computational Models of Argument.* Vol. 305. Warsaw, Poland: IOS Press, pp. 467–468 (cit. on pp. 8, 137).

Meter, Christian, Tobias Krauthoff, and Martin Mauve (July 2017). “discuss: Embedding Dialog-Based Discussions into Websites”. In: *Learning and Collaboration Technologies. Technology in Education.* Vancouver, Canada, pp. 449–460 (cit. on pp. 8, 95, 136).

Meter, Christian, Alexander Schneider, and Martin Mauve (Sept. 2018b). “EDEN: Extensible Discussion Entity Network”. In: *Computational Models of Argument.* Warsaw, pp. 257–268 (cit. on pp. 8, 136).

Modgil, Sanjay and Henry Prakken (2014). “The ASPIC+ framework for structured argumentation: a tutorial”. In: *Argument & Computation* 5.1, pp. 31–62 (cit. on p. 5).

Neumann, Tim, Ole Kelm, and Marco Dohle (2021). “Polarisation and Silencing Others During the Coronavirus Pandemic in Germany: An Experimental Study Using Algorithmically Curated Online Environments”. In: *Javnost – the Public.* Under review (cit. on p. 95).

O'Mahony, Michael P., Neil J. Hurley, and Guénolé C. M. Silvestre (2005). “Recommender Systems: Attack Types and Strategies”. In: *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 1.* AAAI'05. Pittsburgh, Pennsylvania: AAAI Press, pp. 334–339 (cit. on p. 121).

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd (Nov. 1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab (cit. on p. 15).

Pariser, Eli (2011). *The filter bubble: What the Internet is hiding from you.* Penguin UK (cit. on p. 121).

Pazienza, Andrea, Stefano Ferilli, and Floriana Esposito (Nov. 2017a). “Constructing and Evaluating Bipolar Weighted Argumentation Frameworks for Online Debating Systems.” In: *Proceedings of the 1st Workshop on Advances In Argumentation In Artificial Intelligence.*

Ed. by Stefano Bistarelli, Massimiliano Giacomin, and Andrea Pazienza. CEUR Workshop Proceedings 2012. Aachen, pp. 111–125 (cit. on pp. 13, 135).

Pazienza, Andrea, Stefano Ferilli, and Floriana Esposito (2017b). “On the Gradual Acceptability of Arguments in Bipolar Weighted Argumentation Frameworks with Degrees of Trust”. In: *International Symposium on Methodologies for Intelligent Systems*. Springer, pp. 195–204 (cit. on p. 13).

Pollock, John L. (2001). “Defeasible reasoning with variable degrees of justification”. In: *Artificial intelligence* 133.1-2, pp. 233–282 (cit. on p. 5).

Rahman, Md Mahfuzer, Joseph Sirrianni, Xiaoqing (Frank) Liu, and Douglas Adams (2019). “Predicting Opinions Across Multiple Issues in Large Scale Cyber Argumentation Using Collaborative Filtering and Viewpoint Correlation”. In: *The Ninth International Conference on Social Media Technologies, Communication, and Informatics*, pp. 45–51 (cit. on pp. 65, 94).

Romero Moreno, Guillermo, Javier Padilla, and Enrique Chueca (2020). “Learning VAA: A new method for matching users to parties in voting advice applications”. In: *Journal of Elections, Public Opinion and Parties*, pp. 1–19 (cit. on p. 65).

Schafer, J. Ben, Dan Frankowski, Jon Herlocker, and Shilad Sen (2007). “Collaborative Filtering Recommender Systems”. In: *The adaptive web*. Springer, pp. 291–324 (cit. on p. 95).

Schultze, Martin (2012). “Wirkungen des Wahl-O-Mat auf Bürger und Parteien”. In: *MIP 2012*, pp. 127–131 (cit. on pp. 120, 122).

Scott, Alastair J. and George A.F. Seber (1983). “Difference of Proportions From the Same Survey”. In: *The American Statistician* 37.4a, pp. 319–320 (cit. on p. 61).

Selinger, Marcin (2014). “Towards Formal Representation and Evaluation of Arguments”. In: *Argumentation* 28.3, pp. 379–393 (cit. on p. 13).

Silvapulle, Mervyn J. and Pranab Kumar Sen (2011). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. Vol. 912. John Wiley & Sons (cit. on p. 61).

Wagner, Markus and Outi Ruusuvirta (2012). “Matching voters to parties: Voting advice applications and models of party choice”. In: *Acta politica* 47.4, pp. 400–422 (cit. on p. 122).

Walton, Douglas (2009). “Argumentation theory: A very short introduction”. In: *Argumentation in artificial intelligence*. Springer, pp. 1–22 (cit. on p. 5).

Webber, William, Alistair Moffat, and Justin Zobel (2010). “A Similarity Measure for Indefinite Rankings”. In: *ACM Transactions on Information Systems (TOIS)* 28.4, pp. 1–38 (cit. on p. 137).

Xu, Yunwen, Srinivasa M. Salapaka, and Carolyn L. Beck (2013). “A Distance Metric Between Directed Weighted Graphs”. In: *52nd IEEE Conference on Decision and Control*. IEEE, pp. 6359–6364 (cit. on p. 14).

Xuecheng, Liu (1992). “Entropy, distance measure and similarity measure of fuzzy sets and their relations”. In: *Fuzzy sets and systems* 52.3, pp. 305–318 (cit. on p. 65).

Yang, Mei, Sarah Alice Gaggl, and Sebastian Rudolph (2020). “Neva – Extension Visualization for Argumentation Frameworks”. In: *Frontiers in Artificial Intelligence and Applications* 326, pp. 477–478 (cit. on p. 7).

Personal Publications

Conference Papers

Markus Brenneis, Maike Behrendt, Stefan Harmeling, and Martin Mauve (Sept. 2020). "How Much Do I Argue Like You? Towards a Metric on Weighted Argumentation Graphs". In: *Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2020)* (Sept. 8, 2020). Ed. by Sarah A. Gaggl, Matthias Thimm, and Mauro Vallati. CEUR Workshop Proceedings 2672. Aachen, pp. 2–13

Markus Brenneis and Martin Mauve (Nov. 2020b). "Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations". In: *Proceedings of the Workshop on Advances In Argumentation In Artificial Intelligence 2020* (Nov. 25, 2020). Ed. by Bettina Fazzinga, Filippo Furfaro, and Francesco Parisi. CEUR Workshop Proceedings 2777. Aachen, pp. 1–15

Markus Brenneis and Martin Mauve (July 2021b). "ArgVote: Which Party Argues Like Me? Exploring an Argument-Based Voting Advice Application". In: *Intelligent Decision Technologies.* Ed. by I. Czarnowski, R. J. Howlett, and L. C. Jain. Springer Singapore, pp. 3–13

Markus Brenneis and Martin Mauve (July 2021c). "How Intuitive Is It? Comparing Metrics for Attitudes in Argumentation with a Human Baseline". In: *Artificial Intelligence in HCI.* ed. by Helmut Degen and Stavroula Ntoa. Springer International Publishing, pp. 125–138

Markus Brenneis, Maike Behrendt, and Stefan Harmeling (July 2021). "How Will I Argue? A Dataset for Evaluating Recommender Systems for Argumentations". In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue.* Singapore and Online: Association for Computational Linguistics, pp. 360–367

Peer-Reviewed Demo

Markus Brenneis and Martin Mauve (Sept. 2020a). "deliberate – Online Argumentation with Collaborative Filtering". In: *Computational Models of Argument.* Ed. by Henry Prakken, Stefano Bistarelli, Santini Francesco, and Carlo Taticchi. Vol. 326. IOS Press, pp. 453–454

Technical Reports

Markus Brenneis and Martin Mauve (2021a). *ArgVote Bot: Introducing an Argumentative Voting Advice Bot.* Technical Report: Heinrich Heine University Düsseldorf

List of Figures

List of Tables

Eidesstattliche Erklärung
laut §5 der Promotionsordnung vom 15.06.2018

Ich versichere an Eides statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Ort, Datum	Markus Brenneis